Humboldt-Universität zu Berlin

DISSERTATION

A Quantitative Analysis of E-Commerce: Channel Conflicts, Data Mining, and Consumer Privacy

Zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der
Wirtschafswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Herrn Diplom-Ingenieur Maximilian Teltzrow
geboren am 26.03.1975 in Berlin

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jürgen Mlynek

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Joachim Schwalbach

Gutachter: 1. Prof. Oliver Günther, Ph.D.

2. Prof. Alfred Kobsa, Ph.D.

Einreichung der Dissertation: 2. Mai 2005
Datum der Promotion: 17. Mai 2005

Abstract

The role and perception of the Web in its various usage contexts is rapidly changing – from an early focus on "Web-only" interaction with customers, information seekers, and other users, to the Web becoming one central component in a multi-channel information and communication strategy. This development gives companies the opportunity to collect, analyze and use an increasing amount of digital consumer information.

While yielding benefits to the companies (e.g. marketing, usability), the analysis and use of online data has significantly raised consumer privacy concerns, which in turn has become a primary impediment for successful e-commerce. The implications for a company are that it must respect privacy requirements in data analysis and data usage and it must communicate these privacy practices efficiently towards its online visitors.

The aim of this thesis is to explore the border between the competing interests of online consumers and companies. Privacy on the Internet is investigated from a consumer perspective and privacy requirements are specified. A set of business analyses for Web sites is proposed and it is indicated how privacy requirements can be included in the analysis. Moreover, a privacy communication design is presented, which allows more efficient communication of a Web site's privacy practices directed towards the users. The proposed solutions allow the resolution of conflicting goals between companies' data usage and consumers' privacy concerns.

The research is carried out with special emphasis on retailers operating multiple distribution channels. These retailers have become the dominant player in e-commerce. The specific contributions of this thesis are the following:

- *Measuring antecedents of trust in multi-channel retailing (Chapter 2)*

  The success of multi-channel retailing and the importance of privacy is discussed from a consumer's point of view. We present a structural equation model of *consumer trust* in a multi-channel retailer. *Trust* is a well-known predictor of *willingness to buy*.

  A significant influence of *perceived store reputation* and *perceived store size* on *trust* in an e-shop has been identified, which supports our hypothesis that cross-channel effects exist between a retailer's physical store network and its e-shop. We found that consumers' *perceived privacy* had the strongest influence on *trust*. The results suggest to further integrate distribution channels and to improve the communication of privacy online.

- *Design and testing of a Web analysis framework (Chapter 3)*

  Our research on consumer perceptions in multi-channel retailing motivates to further investigate the notion of success measurement on the Internet. We propose an analysis framework consisting of 82 analyses for measuring the online success of Web sites. New conversion success metrics and customer segmentation approaches have been introduced. A particular emphasis has been placed on metrics and analytics for multi-channel retailers. The framework has been tested on Web site data from a large multi-channel retailer and an information site.

- *Prototypical development of a privacy-preserving Web analysis service (Chapter 4)*

  The analysis of Web data requires that privacy restrictions must be adhered to. The impact of privacy requirements on our analysis framework is discussed. We propose a privacy-preserving Web analysis service that calculates the set of 82 business analyses and indicates when an analysis is not compliant with privacy requirements or when data is not available. A syntactical extension of a privacy standard is proposed.

- *Extension of user privacy requirements (Chapter 5)*

  An important application that uses results from the described Web analysis service are personalization systems. These systems become more efficient with an increasing amount of user information. Thus, the impact of privacy concerns is particularly high for personalization applications. An overview of consumer privacy concerns and their particular impact on personalization systems is provided that is summarized in a meta-study of 30 privacy surveys. Approaches to privacy-preserving personalization have been discussed.

- *Development of a privacy communication design (Chapter 6)*

  A company must not only respect privacy requirements in its Web analysis and usage purposes but it must also effectively communicate these privacy practices to its site visitors. A new user interface design approach is proposed, in which the privacy practices of a Web site are explicated in a contextualized manner, and users' benefits in providing personal data clearly explained. A user experiment has been conducted that compared two versions of a personalized store. Subjects who interacted with our new interface design were significantly more willing to share personal data with the Web site. They rated its privacy practices and the perceived benefit significantly higher and made considerably more purchases.

Keywords: Electronic Commerce, Data Mining, Multi-Channel Retailing, Privacy, Communication Design

Zusammenfassung

Die Rolle und Wahrnehmung des World Wide Web in seinen unterschiedlichen Nutzungskontexten ändert sich zunehmend – von einem frühen Fokus auf reine Web-Interaktion mit Kunden, Informationssuchern und anderen Nutzern hin zum Web als eine Komponente in einer mehrkanaligen Informations- und Kommunikationsstrategie. Diese zentrale Entwicklung ermöglicht Firmen, eine wachsende Menge digitaler Konsumenteninformationen zu sammeln, zu analysieren und zu verwerten.

Während Firmen von diesen Daten profitieren (z.B. für Marketingzwecke und zur Verbesserung der Bedienungsfreundlichkeit), hat die Analyse und Nutzung von Onlinedaten zu einem signifikanten Anstieg der Datenschutzbedenken bei Konsumenten geführt, was wiederum ein Haupthindernis für erfolgreichen E-Commerce ist. Die Implikationen für eine Firma sind, dass Datenschutzanforderungen bei der Datenanalyse und -nutzung berücksichtigt und Datenschutzpraktiken effizient nach außen kommuniziert werden müssen.

Diese Dissertation erforscht den Grenzbereich zwischen den scheinbar konkurrierenden Interessen von Onlinekonsumenten und Firmen. Datenschutz im Internet wird aus einer Konsumentenperspektive untersucht und Datenschutzanforderungen werden spezifiziert. Eine Gruppe von Geschäftsanalytiken für Webseiten wird präsentiert und es wird verdeutlicht, wie Datenschutzanforderungen in den Analyseprozess integriert werden können. Darüber hinaus wird ein Design zur effektiven Kommunikation von Datenschutzpraktiken einer Firma gegenüber Konsumenten vorgestellt. Die vorgeschlagenen Lösungsansätze gestatten den beiden Gegenparteien, widerstreitende Interessen zwischen Datennutzung und Datenschutz auszugleichen.

Ein besonderer Fokus dieser Forschungsarbeit liegt auf Mehrkanalhändlern, die den E-Commerce-Markt mittlerweile dominieren. Die Beiträge dieser Arbeit sind im Einzelnen:

- *Messung von Vorbedingungen für Vertrauen im Mehrkanalhandel (Kapitel 2)*

  Der Erfolg des Mehrkanalhandels und die Bedeutung von Datenschutz werden aus einer Konsumentenperspektive dargestellt. Ein Strukturgleichungsmodell zur Erklärung von Konsumentenvertrauen in einen Mehrkanalhändler wird präsentiert. Vertrauen ist wiederum eine zentrale Vorbedingung für die Kaufbereitschaft.

  Ein signifikanter Einfluss der *wahrgenommenen Reputation* und *Größe physischer Filialen* auf das *Vertrauen in einen Onlineshop* wurde festgestellt. Dieses Resultat bestätigt unsere Hypothese, dass kanalübergreifende Effekte zwischen dem physischen Filialnetzwerk und einem Onlineshop existieren. Der *wahrgenommene*

Datenschutz hat im Vergleich den stärksten Einfluss auf das *Vertrauen*. Die Resultate legen nahe, Distributionskanäle weiter zu integrieren und die Kommunikation des Datenschutzes zu verbessern.

- *Design und Test eines Web-Analyse-Systems (Kapitel 3)*

  Der Forschungsbeitrag zu Konsumentenwahrnehmungen im Mehrkanalhandel motiviert die Untersuchung, wie Erfolg im Internet gemessen werden kann. Wir präsentieren ein Kennzahlensystem mit 82 Kennzahlen zur Messung des Onlineerfolges von Webseiten. Neue Konversionsmetriken und Kundensegmentierungsansätze werden vorgestellt. Ein Schwerpunkt liegt auf der Entwicklung von Kennzahlen für Mehrkanalhändler. Das Kennzahlensystem wird auf Daten der Website eines Mehrkanalhändlers und einer Informationswebseite geprüft.

- *Prototypische Entwicklung eines datenschutzwahrenden Web Analyse Services (Kapitel 4)*

  Die Analyse von Webdaten erfordert die Beachtung von Datenschutzrestriktionen. Daher wird der Einfluss von Datenschutzbestimmungen auf das Kennzahlensystem diskutiert. Wir präsentieren einen datenschutzwahrenden Web Analyse Service, der die Kennzahlen unseres Web-Analyse-Systems berechnet und zudem anzeigt, wenn eine Kennzahl im Konflikt mit Datenschutzbestimmungen steht. Eine syntaktische Erweiterung eines etablierten Datenschutzstandards wird vorgeschlagen.

- *Erweiterung der Analyse von Datenschutzbedürfnissen aus Kundensicht (Kapitel 5)*

  Eine wichtige Anwendung, die Resultate des beschriebenen Web Analyse Services nutzt, sind Personalisierungssysteme. Diese Systeme verbessern ihre Effizienz mit zunehmenden Informationen über die Nutzer. Daher sind die Datenschutzbedenken von Webnutzern besonders hoch bei Einsatz dieser Systeme. Datenschutzbedenken aus Konsumentensicht werden in einer Meta-Studie von 30 Datenschutzumfragen kategorisiert, und mögliche Konsequenzen für die Nutzung von Personalisierungssystemen werden beschrieben. Lösungsansätze zur datenschutzwahrenden Personalisierung werden diskutiert.

- *Entwicklung eines Datenschutz-Kommunikationsdesigns (Kapitel 6)*

  Eine Firma muss nicht nur Datenschutzanforderungen bei Web-Analyse- und Datennutzungspraktiken berücksichtigen. Sie muss diese Datenschutzvorkehrungen auch effektiv gegenüber den Seitenbesuchern kommunizieren. Wir präsentieren ein neuartiges Nutzer-Interface-Design, bei dem der Datenschutz und der Kundennutzen der Datenübermittlung auf einer Website klar erläutert wird. Ein Nutzerexperiment wurde durchgeführt, das zwei Versionen eines personalisierten Web-Shops vergleicht.

Teilnehmer, die mit unserem Interface-Design interagierten, waren signifikant häufiger bereit, persönliche Daten mitzuteilen, bewerteten die Datenschutzpraktiken und den Nutzen der Datenpreisgabe höher und kauften wesentlich häufiger.

Schlagworte: Electronic Commerce, Data Mining, Mehrkanalhandel, Datenschutz, Kommunikationsdesign

Declaration

Parts of the structural equation model in Chapter 2 were presented in [Teltzrow, et al., 2003b]. The analysis framework and its empirical evaluation described in Chapter 3 are discussed in [Spiliopoulou, et al., 2002b; Teltzrow and Berendt, 2003; Teltzrow, et al., 2003a; Teltzrow, et al., 2004a; Teltzrow and Günther, 2001; Teltzrow and Günther, 2003]. The integration of privacy impacts and the development of a privacy-preserving analysis prototype in Chapter 4 were discussed in [Boyens, et al., 2002; Teltzrow, et al., 2004b]. The meta-study of privacy surveys in Chapter 5 appeared in [Teltzrow and Kobsa, 2004b]. The experimental study of a privacy communication design in Chapter 6 was discussed in [Berendt and Teltzrow, 2005; Kobsa and Teltzrow, 2004; Kobsa and Teltzrow, 2005; Teltzrow and Kobsa, 2003; Teltzrow and Kobsa, 2004a].

Acknowledgement

There are a number of people without whom this thesis would not have been feasible. Their high academic standards and personal integrity provided me with continuous guidance and support.

Firstly, I would like to thank my doctoral advisor Professor Günther for his continuing support, motivation and valuable input throughout the entire time of this doctoral thesis.

I would like to thank my co-advisor Professor Kobsa for his comprehensive academic guidance and for hosting me at the University of California, Irvine.

My appreciation goes to Professor Berendt who gave me advice, encouragement and valuable feedback innumerable times.

I would like to thank my colleagues from the Berlin-Brandenburg Graduate School in Distributed Information Systems for their inspiring discussions and feedback. I thank the Deutsche Forschungsgemeinschaft for the funding of my scholarship (DFG grant no. GRK 316/3) and the Alexander v. Humboldt Foundation (TransCoop Program) and the National Science Foundation (Grant No. 0308277) for their financial support.

I thank Sören Preibusch for his reliable assistance during this thesis. Thanks also go to my office mates Anett Kralisch and Claus Boyens for their great company.

This thesis is dedicated to my parents, Karin and Peter Teltzrow, without whom none of this would have been possible.

Contents

Figures

Tables

Abbreviations

ABNF	Augmented Backus-Naur Form
AGFI	Adjusted Goodness-of-Fit Index
ASP	Application Service Provider
B2B	Business-to-Business
BDSG	Bundesdatenschutzgesetz [Federal Data Protection Act]
BMP	Bitmap
BWAHLG	Bundeswahlgesetz [German Federal Electoral Law]
CRM	Customer Relationship Management
DNS	Domain Name Server
ECMA	European Computer Manufacturer's Association
EPIC	Electronic Privacy Information Center
EU	European Union
GFI	Goodness-of-Fit Index
GIF	Graphics Interchange Format
IP	Internet Protocol
ISO	International Organization for Standardization
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
KM	Kilometers
P3P	Platform for Privacy Preferences
PGFI	Parsimony Goodness of Fit Index
PNFI	Parsimony Normed Fit Index
PNG	Portable Network Graphics
PCA	Principal Component Analysis
JPEG	Joint Photographic Experts Group

MS	MICROSOFT
PR	PUBLIC RELATIONS
RMSEA	ROOT MEAN SQUARE ERROR OF APPROXIMATION
SQL	STRUCTURED QUERY LANGUAGE
SSL	SECURE SOCKET LAYER
TDDSG	TELEDIENSTEDATENSCHUTZGESETZ [TELESERVICES DATA PROTECTION ACT]
TIFF	TAGGED IMAGE FILE FORMAT
URL	UNIQUE RESOURCE LOCATOR
US	UNITED STATES
WUM	WEB UTILIZATION MINER
W3C	WORLD WIDE WEB CONSORTIUM
WWW	WORLD WIDE WEB
XML	EXTENSIBLE MARKUP LANGUAGE

1 Overview

The role and perception of the Web in its various usage contexts is rapidly changing – from an early focus on "Web-only" interaction with customers, information seekers, and other users, to the Web becoming one central component in a multi-channel information and communication strategy. In fact, multi-channel retailers increased their online market share from 52% in 1999 to 67% in 2001 – in contrast to Internet-only retailers, who lost market share respectively [Silverstein, et al., 2002]. Incumbent companies with a traditional store network seem to dominate the online market currently. With the increasing online competition, measuring success has become crucial for both Web-only and multi-channel retailers.

Web site owners have the opportunity to collect, analyze and use an increasing amount of online consumer information. On the Internet users transmit personal information, either actively by sending customer data (e.g. a shipping address for books), or passively, by leaving traces that are registered with the server side (in the so-called Web server log). In a multi-channel context, Web sites can also collect information about online consumers' use of offline channels. Despite the increasing flow of consumer data, Web sites often lack the ability to utilize the information for measuring e-business success. Multi-channel retailers in particular lack a measurement system to analyze online success in a complex multi-channel information, communication and distribution strategy.

While yielding benefits to the companies, the analysis and use of consumer data increases privacy concerns significantly, which has become a primary impediment for successful e-commerce. Online shoppers claim they would buy considerably more if they were less concerned about their online privacy [Cyber Dialogue, 2001; Department for Trade and Industry, 2001; Forrester, 2001]. Privacy legislation and industry-driven initiatives aim at alleviating some of these concerns. As a consequence, a Web site that aims at analyzing and using online consumer data must include privacy requirements in its analysis practices. Moreover, it must efficiently communicate these privacy standards to its users in order to increase consumer trust.

With regard to Web retailing, we will address the following questions:

- Are there quantifiable cross-channel effects between online and offline retailing that explain why consumers tend to prefer multi-channel over Internet-only retailing?
- Is there a way to assess a Web site's success other than in terms of online purchases?

- How can the notion of online success be measured in a complex multi-channel information, communication, and distribution system?

With regard to online privacy, we will focus on the following questions:

- What are the privacy requirements from a consumer, legal and industry point of view?
- What are potential privacy conflicts between companies' analysis practices and consumers' privacy demands?
- How can these privacy constraints be integrated in a set of structured Web analyses?
- How can a Web site's privacy standards be communicated efficiently to its visitors?

This thesis will propose concrete solutions for the questions raised above.

1.1 Contribution

The thesis' specific contributions are the following:

- *Measuring antecedents of trust in multi-channel retailing (Chapter 2)*

  The success of multi-channel retailing and the importance of privacy is discussed from a consumer perspective. We present a structural equation model of consumer trust in a multi-channel retailer. Trust is a well-known predictor of willingness to buy.

  A significant influence of *perceived store reputation* and *perceived store size* on *trust in an e-shop* has been identified, which supports our hypothesis that cross-channel effects exist between a retailer's physical store network and its e-shop. We found that consumers' *perceived privacy* had the strongest influence on trust. The results suggest to further integrate distribution channels and to improve the communication of privacy online.

- *Design and testing of a Web analysis framework (Chapter 3)*

  Our research on consumer perceptions in multi-channel retailing motivates to further investigate the notion of success measurement on the Internet. We propose an analysis framework consisting of 82 analyses for measuring the online success of Web sites. New conversion success metrics and customer segmentation approaches have been introduced. A particular emphasis has been placed on metrics and analytics for multi-channel retailers. The framework has been tested on Web data from a large multi-channel retailer and an information site.

- *Prototypical development of a privacy-preserving Web analysis service (Chapter 4)*

  The analysis of Web data requires that privacy restrictions must be adhered to. The impact of privacy requirements on our analysis framework is discussed. Legal restrictions and requirements specified in the Platform of Privacy Preferences (P3P)

are presented. We propose a privacy-preserving Web analysis service that calculates the set of 82 business analyses and indicates when an analysis is not compliant with privacy requirements or when data is not available. A syntactical extension of P3P is proposed.

- *Extension of user privacy requirements (Chapter 5)*

  An important application that uses results from the described Web analysis service are personalization systems. These systems become more efficient with an increasing amount of user information. Thus, the impact of privacy concerns is particularly high for personalization applications. An overview of consumer privacy concerns and their particular impact on personalization systems is provided, which is summarized in a meta-study of 30 privacy surveys. Approaches to privacy-preserving personalization have been discussed.

- *Development of a privacy communication design (Chapter 6)*

  A company must not only respect privacy requirements in its Web analysis and usage purposes but it must also effectively communicate these privacy practices to its site visitors. A new user interface design approach is proposed, in which the privacy practices of a Web site are explicated in a contextualized manner, and users' benefits in providing personal data clearly explained. A user experiment has been conducted that compared two versions of a personalized store. Subjects who interacted with our new interface design were significantly more willing to share personal data with the Web site. They rated its privacy practices and the perceived benefit significantly higher and made considerably more purchases.

The thesis concludes with a summary and an outlook on further research in Chapter 7.

A sketch of the thesis structure is captured in Figure 1-1:

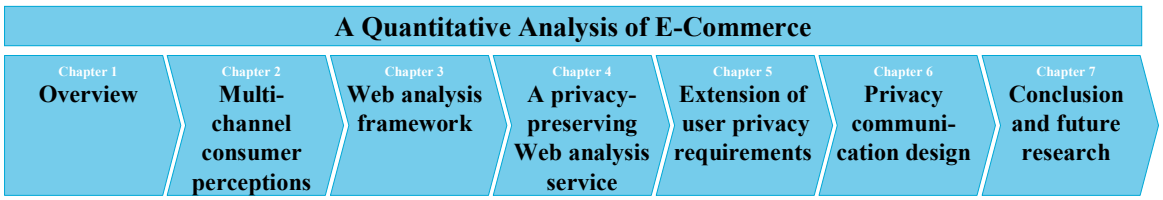

Figure 1-1: Thesis structure

1.2 Methodology of the thesis

This thesis chooses exploratory and confirmatory research approaches that aim at balancing advantages and disadvantages of both theory-building and theory-testing methodologies.

Chapter 2 takes a confirmatory approach to data analysis. Hypotheses are developed and tested on data from 1048 online consumers. Multi-causal relationships have been observed using LISREL 8 [Jöreskog and Sörbom, 2003]. Chapter 3 uses an exploratory research approach. Techniques from data and Web mining are applied on Web user and usage data. The data sample includes customer information from 13,653 customers, 92,467 sessions from a multi-channel retailer's Web logs and external information. Confirmatory elements have been integrated into the Web mining approach in Section 3.7, where background knowledge is used as guidance for the mining process. Preconceptions about the data are tested against a reference set of 27,647 user sessions from a non-commercial site. Chapter 4 develops a prototype, which integrates the exploratory analysis techniques of Chapter 3. Chapter 5 is based on a comparative literature review. Chapter 6 concentrates on an experimental approach. A between-subjects design has been chosen to explore the impact of privacy and personalization communication on users' data disclosure behavior.

2 Multi-channel consumer perceptions

The distribution of products via multiple sales channels — often referred to as multi-channel retailing — is the norm today. According to Silverstein, Sirkin and Stanger [2002] multi-channel retailers in the United States (US) increased their online market share from 52% in 1999 to 67% in 2001 — in contrast to Internet-only retailers, who lost market share respectively. In 2003, multi-channel players comprised 43 of the top 50 e-retailers, versus 42 in 2001, 40 in 2000 and only 27 in 1999 [Gallo and McAlister, 2003]. For some pure Internet retailers a development towards multi-channel retailing can be observed.[1] The increasing prevalence of multi-channel retailing raises the question how the presence of multiple sales channels may influence consumer perceptions of an e-shop and willingness to buy online respectively. In particular, we are interested in whether an effect between the perception of physical stores and *trust* in an e-shop can be measured. *Trust* is an important antecedent of *willingness to buy* [Bhattacherjee, 2002; Gefen, 2000; Koufaris and Hampton-Sosa, 2002; Pavlou, 2003]. Moreover, we are interested in the effect of consumers' *perceived privacy* on *trust* in an e-shop. The results motivate our further research about multi-channel retailing (Chapter 3) and privacy (Chapters 4, 5 and 6).

This chapter is organized as follows. Section 2.1 presents related work. Hypotheses are proposed in Section 2.2 that constitute the basis for the proposed structural equation model. Section 2.3 concentrates on the used methodology. Results are presented in Section 2.4. Section 2.5 discusses the implications and Section 2.6 concludes the chapter with limitations and further work.

2.1 Related work

A number of surveys suggest that the Internet has a distinct influence on offline sales. In a series of studies conducted by the research consultancy Forrester and Shop.org, retailers claimed that about 24% of their offline sales in 2003 were influenced by the Web, which is an increase from 15% in 2002 [Shop.org and Forrester Research, 2004]. A further study estimates that about half of the 60 million consumers in Europe with an Internet connection bought products offline after having investigated prices and details online [Markillie, 2004]. A study by Doyle et al. [2003] analyzed the influence of store perception on online sales. 64.7% of Internet users in 2002 claimed to sometimes or often look at

[1] The largest e-retailer Amazon.com, for example, features products and services from merchants with physical retail stores since 2002.

traditional retail stores and then buy online – up from 50.3% in 2001. The surveys indicate that there are distinct cross-channel effects between online and offline retailing.

Theoretical contributions discuss advantages of multi-channel retailing and demand further empirical work to analyze how the use of multiple channels affects a firm and its customers [Gallaugher, 2002; Goersch, 2003; Gulati and Garino, 2000; Steinfield, 2002; Stone, et al., 2002].

For Internet-only retailers, numerous multivariate models suggest how the perception of certain variables influences consumers' willingness to buy online. A overview of these studies has been provided in Grabner-Kräutner and Kaluscha [2003]. A large number of these studies found that *trust* is one of the most decisive antecedents of consumers' purchase intentions at Internet-only retailers [Grabner-Kräuter and Kaluscha, 2003]. Doney and Cannon [1997] label *trust* even as an order qualifier for purchase decisions. The studies explore a number of antecedents and consequences of consumer trust in online merchants:

Jarvenpaa, Tractinsky and Vitale [2000] developed an Internet trust model that tested the influence of the two independent variables *perceived size* and *perceived reputation* on customers' evaluation of *trust* in a Web site. The model showed that *perceived reputation* had a much stronger effect on *trust* in comparison to *perceived size*. The study was validated in a cross-cultural study by Jarvenpaa [1999] and in a study by Heijden, Verhagen and Creemers [2001]. Moreover, the model suggested that *trust* has a direct influence on *attitude towards the e-shop* and *perceived risk*, which again have an influence on the *willingness to buy*.

Chellappa [2001] hypothesized relationships among the independent variables *perceived privacy* and *perceived security* and the dependent variable *consumer trust* and received significant support in an empirical evaluation. Further aspects of privacy and its influence on trust have been tested by Belanger, Hiller, and Smith [2002]. Recent work has identified privacy as one of the main requirements for successful e-commerce [Ackerman, et al., 1999; Cranor, et al., 1999; Teltzrow and Kobsa, 2004b].

However, none of the reviewed studies explore antecedents of trust in a multi-channel retailer.

2.2 Hypotheses

We are particularly interested in variables influencing *trust* and *willingness to buy* in a multi-channel context. From the described models for Internet-only retailers, we used the repeatedly cross-validated antecedents of *trust*, *perceived reputation* and *perceived size*

as suggested in the literature [Doney and Cannon, 1997; Heijden, et al., 2001; Jarvenpaa, 1999; Jarvenpaa, et al., 2000] to analyze effects on *trust* and *willingness to buy* in a multi-channel setting. In contrast to models dealing with Internet-only retailers, we analyze how *perceived reputation* and *size* of *physical stores* influence trust in an *e-shop*. Moreover, we test the influence of *privacy* on *trust* as proposed in [Chellappa, 2001]. We are particularly interested in the strengths of the relationships when the three antecedents of *trust* – *reputation of stores*, *size of stores* and *privacy* – are measured simultaneously.

As the hypotheses are related to previous studies, we will just briefly discuss the hypotheses of our model and point out our modifications and new research aspects. For a more elaborate discussion of the underlying theory we refer to the original publications.

Jarvenpaa et al. [2000] use the concept of *trust* in the sense of beliefs about trust-relevant characteristics of the Internet merchant. In two empirical studies the authors found support for a significant influence of *perceived size* on *trust* at Internet-only retailers. According to Doney and Cannon [1997] *size* also turned out to significantly influence *trust* in traditional buyer-seller relationships. Large companies indicate existing expertise and resources, which may encourage *trust*. A large store network indicates continuity as stores may not "vanish overnight" [Goersch, 2003]. In a multi-channel context, we assume that the consumer perception of a retailer's physical store presence may also have a positive influence on the perception of consumer trust in the same merchant's e-store. Thus, we hypothesize:

H1: A consumer's trust in an Internet shop is positively related to the perceived size of its store network.

Reputation is defined as the extent to which buyers believe a company is honest and concerned about its customers [Ganesan, 1994]. Consumers may have more trust in a retailer with high reputation because a trustworthy retailer is less likely to jeopardize reputational assets [Jarvenpaa, et al., 2000]. Several empirical studies support the hypothesis that the *reputation of an e-shop* has a strong influence on *consumer trust* in that shop [De Ruyter, et al., 2001; Heijden, et al., 2001; Jarvenpaa, 1999; Jarvenpaa, et al., 2000]. A study of traditional buyer-seller relationships also provided support that *reputation* is an important antecedent of *trust* [Doney and Cannon, 1997]. We assume that the effects observed for a single sales channel may also prove true for the influence of *perceived reputation of physical stores* on *trust* in the same retailer's e-shop.

H2: A consumer's trust in an Internet shop is positively related to the perceived reputation of its store network.

Concerns regarding online privacy have increased considerably and are a major impediment to e-commerce [Teltzrow and Kobsa, 2004b]. Consumer privacy concerns are particularly elevated on the Internet. A measurement scale for *perceived privacy* towards an e-shop has been suggested by Chellappa (2001) where privacy has been described as the anticipation of how data is collected and used by a marketer. The author also found empirical support that *perceived online privacy* towards an e-shop is significantly related to *consumer trust*. We are interested in replicating this effect in a multi-channel setting.

H3: A consumer's trust in an e-shop is positively related to the perceived privacy at the e-shop.

Trust is closely related to *risk* [Hawes, et al., 1989]. Jarvenpaa et al. [2000] refer to *risk perception* as the "trustor's belief about likelihoods of gains and losses". The hypothesis has been confirmed that the more people trust an e-shop, the lower the perceived risk perception [Heijden, et al., 2001; Jarvenpaa, 1999; Jarvenpaa, et al., 2000]. We also test this hypothesis in our study.

H4: Perceived risk at an e-shop is negatively influenced by consumer trust in an e-shop.

The theory of planned behavior suggests that a consumer is more willing to buy from an Internet store which is perceived as low risk [Ajzen, 1991]. The trust-oriented model by Jarvenpaa (2000) suggests that consumers' *willingness to buy* is influenced by *perceived risk* and *attitude towards an e-shop*. In the studies of Bhattacherjee [2002] and Gefen [2000], a direct influence of *trust* on *willingness to buy* has been suggested. However, Bhattacherjee [2002] states that a large proportion of variance was left unexplained, which suggests that there may be other predictors that are missing in the model. We analyzed the causal relationships between *risk*, and *purchase intention* tested by Jarvenpaa et al. [2000].

H5: The lower the consumer's perceived risk associated with buying from an e-shop, the more favorable are the consumer's purchase intentions towards shopping at that e-shop.

2.3 Methodology

We introduce the methodical approach to test the above hypotheses. The retailer, the questionnaire, respondents' demographics and the statistical method to develop our model are presented.

2.3.1 The retailer

The above hypotheses will be tested using a survey of visitors of a large multi-channel retail Web site. The company's retail site considers itself the first fully integrated multi-

channel shop in Europe. The retailer operates an e-shop and a network of more than 6,000 stores in over 10 European countries. The company sells more than 10,000 consumer electronics products both online and offline. The offered product assortment appeals to a variety of consumer types including bargain shoppers and quality-oriented high-end buyers.

The retail site uses a typical online privacy statement that can be accessed through a link on each page of the site.

A questionnaire could be accessed via a rotating banner on the retail site. The banner announcing the survey was kept online for five months from 1st of March to 30th of July 2004. The banner announced the survey and offered an optional raffle (cf. Figure 0-1 of the Appendix). All participants who left their e-mail address automatically participated in the raffle of three digital cameras.

2.3.2 Questionnaire

The answers on the online questionnaire were measured using a Likert scale ranging from 1 to 5, with 1 indicating an attribute was “very weak / unlikely” and 5 “very strong / likely” [Likert, 1932]. Demographic information included age, gender, Internet experience, e-mail address and questions about previous visits and purchases both online and offline.

Scales were constructed on the basis of past literature as shown in Table 0-1 of the Appendix. For each item of the constructs *perceived size* and *perceived reputation* the term “this Web site” was replaced with “this retailer’s physical store network” to emphasize the offline context. For the remaining items we used the term “this e-shop” to draw a clear distinction between online and offline presence.

Some modifications of the scale suggested by Jarvenpaa [1999; 2000] were adapted from Heijden et al. [2001]. For the construct *willingness to buy*, we changed the time horizons “three months” and “the next year” to the broader terms “short term” and “the longer term”. For each construct we used only three items to keep the questionnaire as short as possible, which was a requirement from the cooperation partner. We also modified two items of the risk scale suggested by Jarvenpaa [1999; 2000] after a pre-test with department faculty. The item “How would you characterize the decision to buy a product through this Web site?” with answers ranging from “a very negative situation” to “a very positive situation” was changed into “How would you characterize the risk to purchase at this e-shop?” with a scale ranging from “very low risk” to “very high risk”. We also introduced a new item to measure consumer perceptions of the store network size: “This retailer’s stores are spread all over the country”. Members of the faculty staff and students

reviewed a preliminary version of the measurement instrument with respect to precision and clearness. A pre-test of 30 participants showed satisfactory results for Cronbach's Alphas [cf. Cronbach, 1951].

2.3.3 Pre-processing and respondents' demographics

Records of 266 respondents were eliminated from a total of 1314 due to missing data, in which duplicated e-mail addresses occurred (41 entries) or text fields belonged apparently to the same participant. 1048 complete answer sets are used for modeling.

The user demographics of our sample is predominantly male and between 30-50 years old (cf. Figure 2-1). 73% of the respondents in our sample are male (n=770) and 26% female (n=278). Thus, it reflects the gender gap that still predominates Internet usage in Europe [Hupprich and Fan, 2004]. Most of the users in our sample are experienced in using the Internet (cf. Figure 2-2).

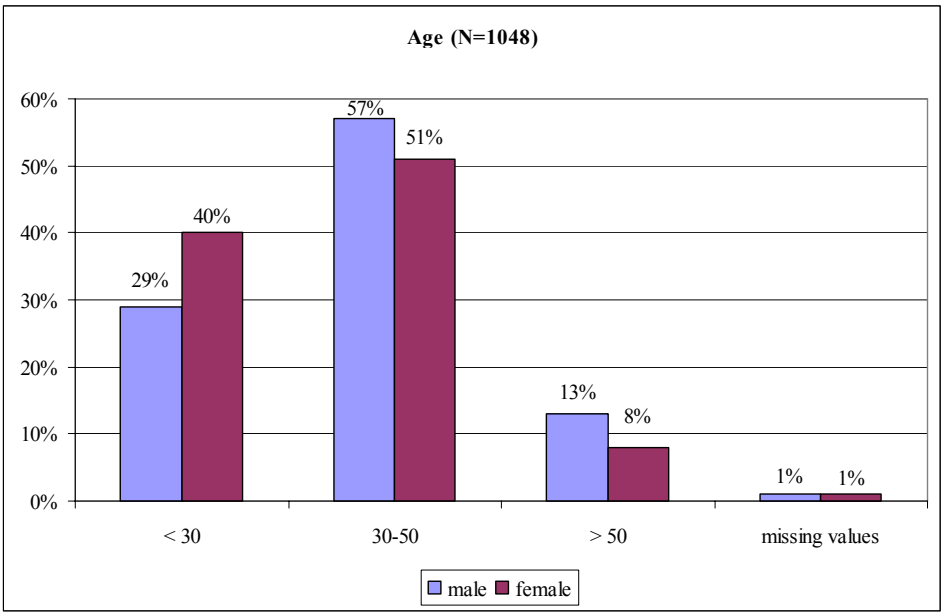

Figure 2-1: Age distribution in respondent sample

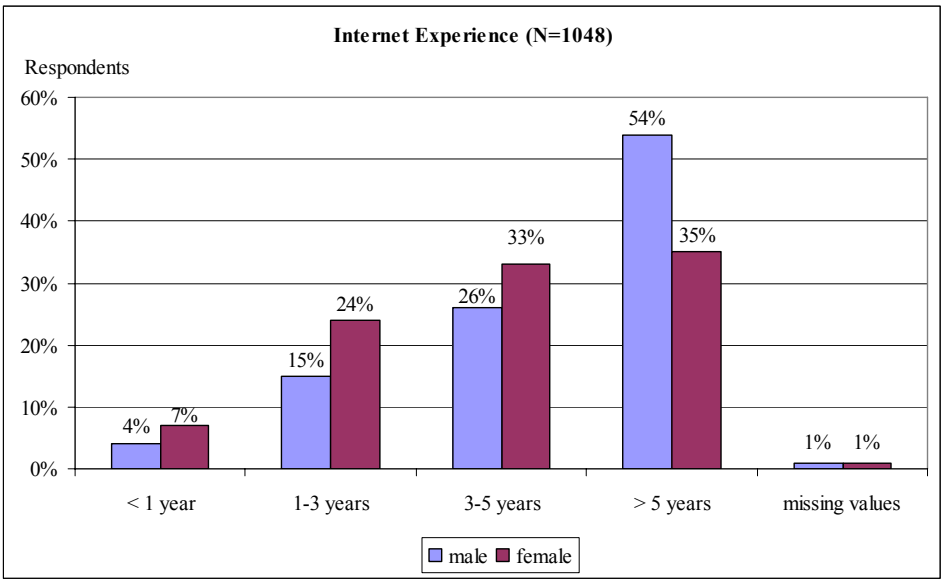

Figure 2-2: Internet experience in respondent sample

Moreover, participants were asked in the questionnaire about their channel experience prior to their actual visit. For each of the four incidents “purchased at e-shop”, “purchased at store”, “visited e-shop” and “visited store”, participants were asked to answer if and how often they had visited the e-shop or store and if and how often they had purchased in the e-shop or in-store. The answers are depicted in Table 2-1.

	visited e-shop	visited store	purchased at e-shop	purchased at store
no visit	300	337	818	425
1-2 times	243	274	168	320
3-5 times	101	111	26	85
> 5 times	388	315	20	200
no answer	16	11	16	18
Total	1048	1048	1048	1048

Table 2-1: Prior experiences with the retailer’s e-shop and stores

It is interesting that more than 605 participants claimed they had purchased at least once at the store and only 214 claimed to have purchased at the e-shop. Moreover, 200 claimed that they had purchased more than five times at a retail store. In contrast, the number of people who visited the store at least once was almost equal to the number of visitors who visited the e-shop at least once. These numbers remarkably point out the

importance of physical stores to the online audience.

2.3.4 Factor analysis and structural modeling

We used cross-validation and divided the sample of 1048 records into two sub-samples $n_1 = n_2 = 524$ using simple random sampling. A confirmatory factor analysis (oblimin rotation) [Jennrich and Sampson, 1966] is performed on sample 1. This analysis was intended to confirm the hypothesized scales in terms of the discovery of six factors that each make up the employed scales.

If a plausible factor structure could be identified, it would be desirable to quantify the effect of *perceived size*, *perceived reputation of stores*, and *privacy* onto *trust*, *willingness to buy*, and *risk perception*. Factors are latent (not directly observable) variables. Linear structural modeling is used here as it allows the simultaneous mapping of relationships between several latent and non-observable variables within a single multi-equation model [Jöreskog and Sörbom, 1979; Jöreskog and Sörbom, 1996].

The variables of the questionnaire have ordinal scales. Model specification and parameter estimation is based on SIMPLIS [Jöreskog and Sörbom, 1996] and LISREL 8.54 [Jöreskog and Sörbom, 1996], and uses only sample 1 units. The model parameters are estimated by weighted least squares algorithm [Jöreskog and Sörbom, 1996]. Model structures were learned and the parameter estimated in an explorative and iterative way. Then the induced model is tested in the following phase on sample 2 in order to guarantee unbiased measures of goodness of fit.

2.4 Results

Firstly, we present a factor analysis, secondly evidences derived from the model, and finally we close with remarks on privacy and trust of respondents.

2.4.1 Factor analysis

The factor analysis included all items from Table 0-1 of the Appendix. The “eigenvalue > 1” - criterion leads to an initial five-factor model. However, a strong evident decline in the scree-plot after the sixth factor demanded a rotation with six factors. The extraction with principal component analysis (PCA), and oblimin rotation (delta = 0°) resulted in 74% explained variance. The first factor has a relatively high fraction of the overall variance, i.e. 33.9%. After rotation, all factors had eigenvalues above 2.

Four factors displayed medium intercorrelations (see Table 2-2), which underlines the necessity of an (oblimin) rotation. The pattern matrix of the rotated solution can be found in Table 0-2 and the factor loading in Table 0-1 of the Appendix.

	I	II	III	IV	V	VI
I	1.00	.02	.31	.42	.37	-.39
II	.02	1.00	-.08	.07	.12	-.06
III	.31	-.08	1.00	.25	.20	-.27
IV	.42	.07	.25	1.00	.19	-.25
V	.37	.12	.20	.19	1.00	-.19
VI	-.39	-.06	-.27	-.25	-.19	1.00

Table 2-2: Factor inter-correlation matrix

All factors include three items each with high factor loading above .6, except for the last factor, cf. -.52, -.58 and -.76. All items that load a factor have the same scale. The factors allow testing of models of causal influence between factors based on linear structural modeling. The medium factor correlation between factors I and III, I and IV, I and V, and I and VI already indicate that such influences exist.

2.4.2 Linear structural models

To test our main five hypotheses, the six factors identified above are inserted into a linear structural model as proposed in Section 2.2. Linear structural models allow the testing of hypotheses about causal influences between latent (not directly observable) variables. As factors, as identified in the previous section, are latent variables (constructs that influence groups of items), hypotheses about their influence on each other can be tested. In linear structural models, factors are displayed as circles. The items that are influenced by these factors are displayed as boxes. Causal influences are displayed as pointed arrows with path coefficients (between -1 and 1) that indicate the strength of the causal relation. Correlations are displayed as bi-directional arrows. By quantifying the influence of the factors on the items, the model may confirm the factor analysis from the previous section.

The models were developed with the SIMPLIS command language [Jöreskog and Sörbom, 1996] and LISREL 8.54 [Jöreskog and Sörbom, 2003]. Due to the fact that ordinal questionnaire data was used, the weighted least squares algorithm for polychoric correlations was employed, including the asymptotic covariance matrices [Jöreskog and Sörbom, 1996].

However, stable parameter estimates of the model could not be determined after 30 iterations. Consequently, the model is reduced to a simpler one, which tried to capture the

effect of different factors on *trust*. A model is iteratively searched, which includes the factors *perceived size* (PS), *perceived reputation* (PR), and *privacy* (PRI). The underlying assumption of this model is that these three factors determine *trust* (TR). This model produced stable parameter estimates and after incorporating a series of modification indices supplied by the LISREL software, reached optimal fit indices. The completed model for sample 1 with all standard errors, factor loadings, and path coefficients is depicted in Figure 2-3.

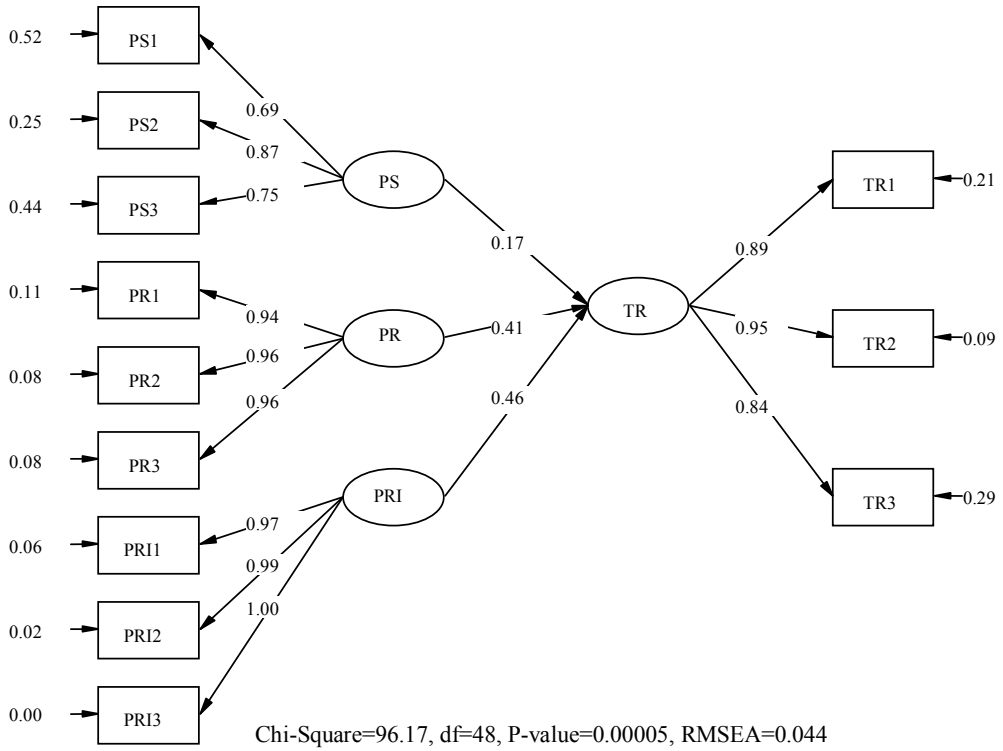

Figure 2-3: Linear structural model for the influence of perceived size (PS), perceived reputation (PR), privacy (PRI) on trust (TR) for sample 1 (N=524)

All path coefficients are significant on the 5% level using a t-test. Goodness of fit statistics gives a Chi square value of 96.17 with 48 degrees of freedom, leading to a p-value of 0.00005[2]. Since the Chi square fit index in linear structural models is highly dependent on the sample size [Byrne, 1998] and tends to underestimate the model fit in larger samples, further fit indices are considered for model assessment. The Root Mean Square Error of Approximation (RMSEA) of 0.044 leads to a p-value for Test of Close Fit of .778, which

[2] Note that in linear structural models, the model hypothesis is that the empirical parameter matrix and the model matrix are not different, thus the p-value has to be as high as possible and not below 0.05.

indicates a good model fit. A Goodness-of-Fit Index (GFI) of 0.99, an Adjusted Goodness-of-Fit Index (AGFI) of 0.99, a Parsimony Normed Fit Index (PNFI) of 0.721 and a Parsimony Goodness of Fit Index (PGFI) of 0.612 supports a good overall model fit. Refer to Jöreskog and Sörbom [2003] for detailed information on fit indices.

These above measures may be biased since the model is induced from the same sample. An unbiased test of the model can be achieved by applying it to the second sample that remained untouched so far (see Figure 2-4).

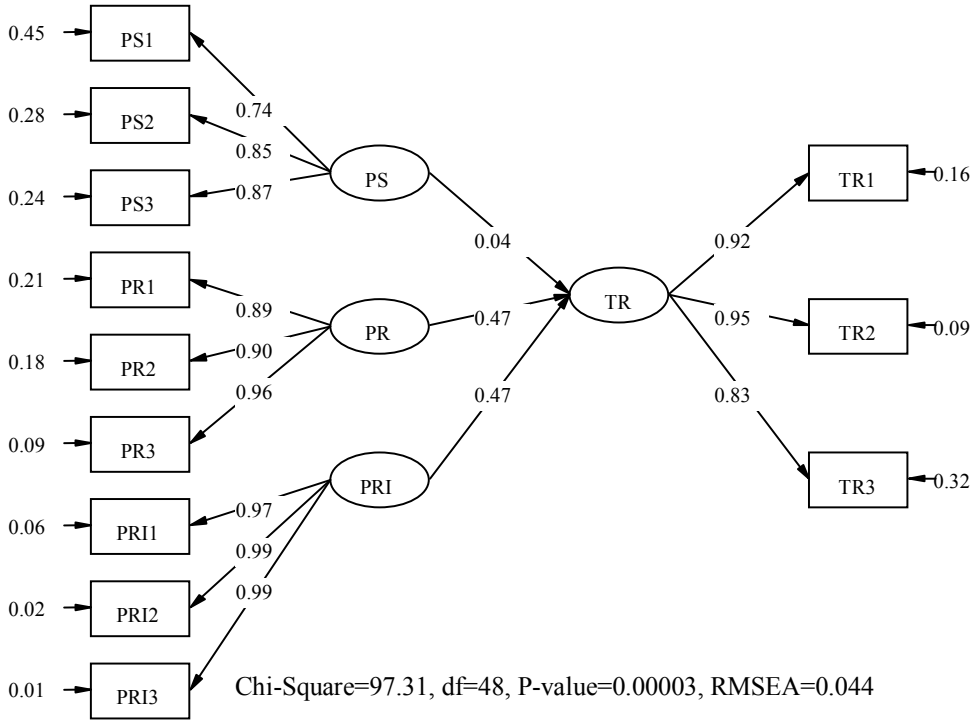

Figure 2-4: Linear structural model for the influence of perceived size (PS), perceived reputation (PR), privacy (PRI) on trust (TR) for sample 2 (N=524)

The model for sample 2 gives a Chi-square value of 97.31 with 48 degrees of freedom, leading to a p-value of 0.00003. This RMSEA-value of 0.044 leads to a p-value for Test of Close Fit of .758, a PGFI of .611, a PNFI of .719 and an AGFI of .996. In summary, these measures point out a good model fit with path coefficients in the same range as in the previous model, cf. Figure 2-3. The relevant path coefficients and fit indices for the two sub-samples as well as for the full sample are summarized in Table 2-3. All path coefficients in the samples are significant on the 5% level except the coefficient PS→ TR in the second sub sample. However, the coefficient is significant in the full sample.

Sample	N	Path PS➔TR	Path PR➔TR	Path PRI➔TR	X^2	df	*P*	RMSEA	*P* (Cl. Fit)
1st	524	0.17*	0.41*	0.46*	96.17	48	0.00005	0.044	0.778
2nd	524	0.04*	0.47*	0.47*	97.31	48	0.00003	0.044	0.758
Full	1048	0.11*	0.42*	0.46*	106.80	48	0	0.034	0.999

Table 2-3: Relevant path coefficients and fit indices for sub samples and entire sample

With regard to Section 2.2, the findings support hypotheses 1-3. Hypothesis 4 assuming a negative influence of *trust* on *risk* and hypothesis 5 assuming an influence of *perceived risk* on *trust* have not been fully confirmed with the conservative methodical approach presented above. Further work will analyze the mediation path between *trust*, *risk* and *willingness to buy* in more detail.

2.5 Discussion and implications

The results indicate that *perceived online privacy* has the highest influence on *trust* relative to the two variables *perceived size of the store* network and *reputation of the store network*. This result has been confirmed in two random samples each with a high P-value. Though surveys indicate that *privacy* is crucial to successful e-commerce [Teltzrow and Kobsa, 2004b], very few of the monthly site visitors accessed the retailer's privacy statement, which is a typical phenomenon at retail sites. Kohavi [2001] indicates that less than 0.5% of all users read privacy policies. As a consequence, retailers should place clear and readily available privacy explanations on their Web sites in order to increase consumer trust. An efficient privacy communication design will be discussed in Chapter 6.

Moreover, the results confirm a strong effect of *perceived store reputation* on *trust* in the e-shop. A small effect of *perceived store size* on *trust* is observed. Thus, our study confirms the existence of cross-channel effects between stores and Web site. Jarvenpaa [2000] has shown that *reputation* and *size* are important antecedents of *trust* at Internet-only retailers. Her speculation that the presence of physical stores might increase consumers' trust in a seller's Internet store can be supported with our results. It can be assumed that cumulative effects between consumers' perceptions of *online* and *offline reputation* and *size* exist. This could be an explanation as to why consumers prefer multi-channel retailers that now dominate more than two-thirds of the total online market (Silverstein et al. 2002). Thus, retailers' multi-channel strategies should increasingly promote trust-building measures between different sales channels. This could include in-store advertising of the Web site, detailed online information about offline stores, better

multi-channel service integration or the placement of in-store kiosks, where consumers can order online when products are out-of-stock. Further studies should explore if there are cumulative effects between the *perceived reputation* and *size of the e-shop* on *trust in the e-shop* as indicated by Heijden et al. [2001] and Jarvenpaa [1999; 2000] and the observed influence of *perceived store size* and *reputation* on *trust in the e-shop*. Therefore, a larger sample of consumers is required for discriminating between three groups of visitors: “familiar with the Web site only”, “familiar with stores only”, and “familiar with both channels”.

An interesting improvement of our study is a further analysis of the variables *trust*, *risk* and *willingness to buy*. Several authors have suggested a direct influence of *trust* on *willingness to buy* on the Internet [Bhattacherjee, 2002; Gefen, 2000; Koufaris and Hampton-Sosa, 2002; Pavlou, 2003]. The relationship between *trust* and success of relationship marketing is also well-known in traditional marketing theory [Berry, 1995; Morgan and Hunt, 1994]. In further work we will test if the construct *perceived risk* may function as a mediator between *trust* and *willingness to buy*. A mediator hypothesis between *trust* and *future intentions* also has been suggested in Garbarino and Johnson [1999]. The authors found that a model where *satisfaction* has been added as a mediating path between *trust* and *commitment* significantly improves the model fit compared to a model suggesting a direct influence of *trust* on *future intentions*.

2.6 Limitations

Participants in this study were online consumers. Thus, the sample differs positively from many other empirical studies that primarily use students as a sample of online consumer population [Grabner-Kräuter and Kaluscha, 2003]. However, a limit of external validity within our sample could have occurred through the self-selection of online participants. Other problems of online questionnaires could be reduced: repeated entries could be widely eliminated as most participants provided demographic information and e-mail addresses to participate in the raffle. The use of a rotating banner added randomness to the selection of participants. Only about every sixth visitor saw the banner on the retailer’s home page. Moreover, we explicitly asked participants to provide only honest answers.

The types of products may influence a user’s willingness to buy [Jarvenpaa, et al., 2000], which has not been further considered in this study. The results of Jarvenpaa et al. suggest that *perceived size* and *reputation* may influence *trust* differently depending on the type of products offered. The product sector of consumer electronics tends to be highly suitable for multi-channel retailing [Omwando, 2002]. It could be that the observed effects are less significant for less Internet-suitable product portfolios. A deeper discussion

of product characteristics in multi-channel retailing can be found in the thesis by Goersch [2003]. Critique also concerns the definition of measurement scales [Grabner-Kräuter and Kaluscha, 2003]. We used scales that have been successfully applied in studies of Internet-only retailing. The scales included relatively few items per construct due to the retailer's request to keep our survey as short as possible. Though the results returned good factor confirmation scores, scaling needs more attention in further studies.

“A science is as mature as its measurement tools.” (Louis Pasteur)

3 Web analysis framework

Chapter 2 highlighted success factors of multi-channel retailing and emphasized the importance of privacy protection on the Internet. The results motivate our further work on success measurement in Web retailing and on the protection of consumer privacy.

This chapter introduces an analysis framework for measuring online success on multi-channel and Internet-only sites. Our analysis framework will propose five categories of business analyses that aim at measuring notions of online success.

The analysis results are particularly useful for customer relationship management and personalization, which will be discussed in more detail in Chapter 5.

This chapter is organized as follows. Section 3.1 presents the data used to test our analysis framework. Section 3.2 introduces a terminology of business analyses and presents the five analysis categories that constitute our analysis framework. Section 3.3 presents a set of service analyses for Web sites of multi-channel retailers. Based on a systematic distinction of service options in multi-channel and Internet-only retailing, we derive analyses measuring online consumers’ service preferences in multi-channel retailing. Section 3.4 proposes a set of Web analyses measuring conversion success. We formalize existing conversion metrics that have so far been described only informally. New metrics are proposed measuring conversion success in a multi-channel context. Section 3.5 extends the analysis of purchase sessions by using a clustering approach, which provides detailed insight into customers’ usage patterns. The analysis is based on a combination of Web usage and Web user data. Section 3.6 presents analyses for consumer segmentation based on demographic and order data. Section 3.6 proposes segmentation approaches indicating a customer’s value to a company. Concentration indices are introduced and an index of customer value is presented. Section 3.7 presents an approach how success can be measured on information Web sites. A mining template for modeling behavioral strategies as sequences of tasks is introduced.

The proposed analyses of Sections 4.3-4.6 are applied to Web user and usage data from the multi-channel retailer presented in Section 2.3.1. Results of Section 3.7 are presented based on data from an information Web site.

3.1 Data

This section presents the data used for the empirical testing of the Web analysis framework for Internet-only and multi-channel Web sites.

Web site owners can collect two types of consumer data: actively divulged Web *user* data and passively transmitted Web *usage* data. Consumers actively divulge user data when they send information to a Web site for billing purposes, to register or request information. Visitors passively transmit usage data by leaving traces registered with the Web site server.

Data from two Web sites have been used to exemplarily calculate the proposed metrics and analytics in our Web analysis framework. Based on the multi-channel retailer introduced in Section 2.3.1 we analyzed 92,467 sessions taken from a period of 21 days in 2002, and transaction information of 13,653 customers who conducted 14,957 online purchases over a period of 8 months in 2001/02. From an information Web site, we analyzed a reference set of 27,647 user sessions.

Section 3.1.1 and 3.1.2 present the structure and terminology of Web user and usage data. The data model of the multi-channel retailer is also presented.

3.1.1 Web usage data

Server logging is based on a protocol component that registers requests to a World Wide Web (WWW) server. These server requests can be initiated by a *user* who visits a *Web site* consisting of many *Web pages*. Each *Web page* is composed of constituent objects such as body text, images or video files, which count as a *hit* each when invoked. Thus, each page a user views (*page view*) comprises many hits at the server. A *clickstream* is a time-ordered list of page views. A *user session* is a set of users' server requests to one or more Web servers. *Sessions* are also referred to as *visits* [Monticino, 1998].

A standard format for logging server requests has been established by the World Wide Web Consortium [W3C, 1995].

The following log entry, taken from the multi-channel retailer's Web server, exemplifies the main parts of an access log in the Extended Log File Format.

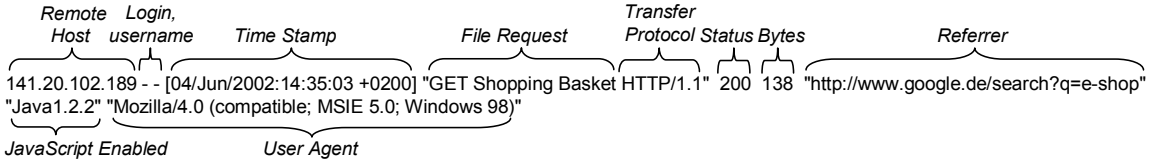

Figure 3-1: Simplified log entry from the cooperation partner

The first part of the log file is the remote host address (Internet Protocol (IP) address), which can be used to identify a visitor's computer or device. The IP address is a 32 bit-long, dotted decimal notation, in which each byte is shown as a decimal number encoded

by 8 bits. It can be translated to a domain name via the Domain Name Server (DNS). The first part of the IP address identifies the user's network address (e.g. 141.20 is the network of the Computer Science department at Humboldt Universität zu Berlin) and may reveal information about the network owner. The last two digits of the IP address specify the host (end-system) within the network, which are assigned (uniquely or dynamically) to a computer.

The DNS can be used to determine a user's geographic location [Lamm, et al., 1996]. Software vendors claim that they can link IP addresses to geographic locations with an accuracy of 98% for *country*, 70% for *regional*, and 65% for *city* level [Melissa Data, 2004; Olsen, 2000]. A source of inaccuracy for geographic localization is the use of proxy servers, which only reveal the location of the proxy server but not the location of the user.

The log file also contains the *remote login name* and *user authentication* of the user if the site requires logins to access a Web server. Moreover, it contains the date and *time of a user request*, the *file name* (e.g. of a Web page, picture, document), the *number of bytes transferred* and the method the client used to retrieve a file from the server (typically GET). The HyperText Transfer Protocol (HTTP) response code (*status code*) indicates the success or failure of the file transfer. The *referrer* indicates the Unique Resource Locator (URL) of the previous page request and the *user agent* indicates browser type and version the client claims to be using. If a site offers active program components, information about a user's JavaScript availability, installed plug-ins or screen resolution can be collected.

When the user leaves name, address or other identifying information on a Web site (e.g. in registration or purchase forms) a unique identification can be assigned to the log file to combine personal information and the respective clickstream.

*Session identifications (session id*s), *cookies* or *IP addresses* can be used to identify and reconstruct a user session. The process of reconstructing the activity log into sessions is referred to as *sessionizing* [Berendt, et al., 2001].

Cookies are small text files stored on a user's hard drive and can be used to recognize users in later sessions. *Session ids* can be transient cookies that are only stored temporarily during a single session and are embedded in the URL. However, users can delete cookies. A recent study claims that 55% of all cookies become unusable each month [Fiutak, 2004]. Further, the use of cookies can have privacy implications, which will be discussed in Section 5.3.

Table 3-1 shows a simplified session sample from the multi-channel retailer's Web site. Sessions were determined by the use of session ids, which are available in the log file.

```
141.20.102.189 - - [04/Jun/2002:14:36:12 +0200] "GET  Home HTTP/1.0 SessionID
bhApYI6N"  200  6500  "http://www.google.de/search?q=e-shop"  "Java1.2.2"
"Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

141.20.102.189 - - [04/Jun/2002:14:37:24 +0200] "GET Browse_Catalog Catalog ID
7n66hz3 HTTP/1.0 SessionID bhApYI6N" 200 759 "http://www.e-shop.de/home"
"Java1.2.2" "Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"

141.20.102.189 - - [04/Jun/2002:14:42:54 +0200] "GET View_Product Product ID
19453 HTTP/1.0 SessionID bhApYI6N " 200 759 "http://www.e-shop.de/Browse_Catalog
CatalogID 7n66hz3" "Java1.2.2" "Mozilla/4.0 (compatible; MSIE 5.5; Windows NT
5.0)"

141.20.102.189  -  -  [04/Jun/2002:14:53:21  +0200]  "GET  BasketForm
PaymentTransactionID  3dNC4KHg  PlacedOrderID  3d4rEKHgFoT  http://www.e-
shop.de/ViewBasket PaymentTransactionID 3dNC4KHg HTTP/1.0 SessionID bhApYI6N"
200 7258 "-" "Java1.2.2" "Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0)"
```

Table 3-1: Session sample from the multi-channel retailer

The time stamps between subsequent page requests can be used to derive users' *view times* per page. Some requested files of the multi-channel retailer contain a `catalog_id` indicating a specific catalog category, a `product_id` indicating a product, a `transaction_id` representing the invocation of the transaction phase, an `order_id` denouncing a purchase and a `payment_id` indicating the chosen payment method.

Before the data is stored for analysis purposes, the typical data cleaning steps in Web mining such as robot removal need to be performed. We abstained from analyzing page view times as reconstructing view times is subject to significant inaccuracies [Berendt, et al., 2001].

Several technical problems may complicate the use and processing of log files. In particular, *caching*, the use of *proxy servers*, *dynamic IP addresses* and the *use of a device by several people* pose a challenge to session reconstruction and user identification [Berendt, et al., 2001; Büchner, et al., 1999; Cooley, et al., 1999; Spiliopoulou, et al., 2003; Wilde, 2003].

For the analysis of user behavior it is beneficial to codify page requests as *session vectors*. Given a set of n pageviews, $P = \{p_1, p_2, \ldots, p_n\}$, and a set of m user session, $S = \{s_1, s_2, \ldots, s_m\}$, where each $s_i \in S$ is a subset of P, each user session can be regarded as a vector over the n-dimensional space of pageviews. The session vector is given by: $\vec{v} = (w^s_{p1}, w^s_{p2}, \ldots, w^s_{pn})$, where $w(p^s_i)$ is the weight associated with pageview p^t_i in the session s_i representing its significance. Usually, but not exclusively, the weight is

based on number of pages visited or page view time, where each $w^s_{pj} = w(p^s_i)$, for some $i \in \{1,\ldots, n\}$, in case p_j appears in the session s_i, and otherwise $w^s_{pj} = 0$. [Dai and Mobasher, 2003]. Thus, conceptually, the set of all user sessions can be viewed as an m × n session pageview matrix.

3.1.2 Web user data

The multi-channel retailer uses a relational database schema to store billing information. The following list represents a simplified view on the retailer's data schema including preprocessed and sessionized Web log data (cf. table **session)**. The company's full database consists of more than 30 tables and 200 attributes. The following list presents those entities and relationships that were used to test the main parts of our analysis framework in Chapter 3 and for the discussion of privacy problems in Chapter 4.

customer (customer_id, geo_id, credit_rating, first_name, surname, title, gender, date_of_birth)

address (address_id, customer_id, geo_id, country_code, street, street_number, street_number_supplement, customer_zip_code, town, recipient_address, post_office_box, phone_number, e-mail_address)

order (order_id, customer_id, session_id, store_id, product_id status, invoice_value, currency, order_date, order_time, delivery_type, payment_method, credit_card_no, customer_card_no, status_change)

product (product_id, category_id, product_name, product_weight, product_size, price, cost)

product_category (category_id, category_name)

return (return_id, order_id, store_id, return_date, return_value, return_address)

store (store_id, geo_id, store_country_code, store_street_name, store_street_number, store_zip_code, store_town)

session (session_id, order_id, ip_location, access_time, browser_type, status_code, referrer)

page (page_id, concept_id, session_id, page_name, page_content)

page_concept (concept_id, concept_name, concept_content)

belongs_to (page_id, concept_id)

contains (session_id, page_id)

location_zip (geo id, micro id, zip_code, longitude_zip_code, latitude_zip_code)
microgeography (micro id, detail_type, detail_value)
characterizes (micro id, geo id)

Table 3-2: User data schema

Foreign keys establish relationships between tables and are depicted as dotted attributes in the presented data schema. Log data in the table **session** could be linked to attributes in the table **customer** via a unique order_id when a user made an online purchase. If a site uses cookies, the attribute cookie_id would be stored in the table **session**.

Third-party data sources can be added to extend a retailer's database with additional consumer profile information. We acquired demographic data from Deutsche Post Direkt [Deutsche Post Direkt GmbH, 2004] that matches zip codes and geographic coordinates. Thus, the table **location_zip** could be added.

Demographic and sociographic information can be linked to customer addresses. The column detail_type in the table **microgeography** includes the attributes that could be added via the geo_id (e.g. zip code). Data such as political orientation, car type, family structure, cultural background, status, spending capacity, household size, creditworthiness, age, preferred anonymity level, marketing affinity, product affinity, preferred order medium or preferred communication media can be purchased from external sources [Deutsche Post Direkt GmbH, 2004]. Due to changes in demography and lifestyles the accuracy and timeliness of microgeographic data is limited, however [Weichert, 2004].

Multi-channel retailers can link data from other sales channels to further enrich customer data. For example, shopping cards[3] are often used to collect and link data from multiple sales points and may allow the detailed tracking of a customer's shopping history.

The entity-relationship model for the multi-channel retailer is depicted in Figure 3-2.

[3] With shopping cards customers can earn bonus points for each purchase, which can be redeemed in the form of discounts and/or other incentives. Though data from shopping cards is valuable for marketing, there is a potential bias because cardholders may have a stronger brand loyalty than the average customer.

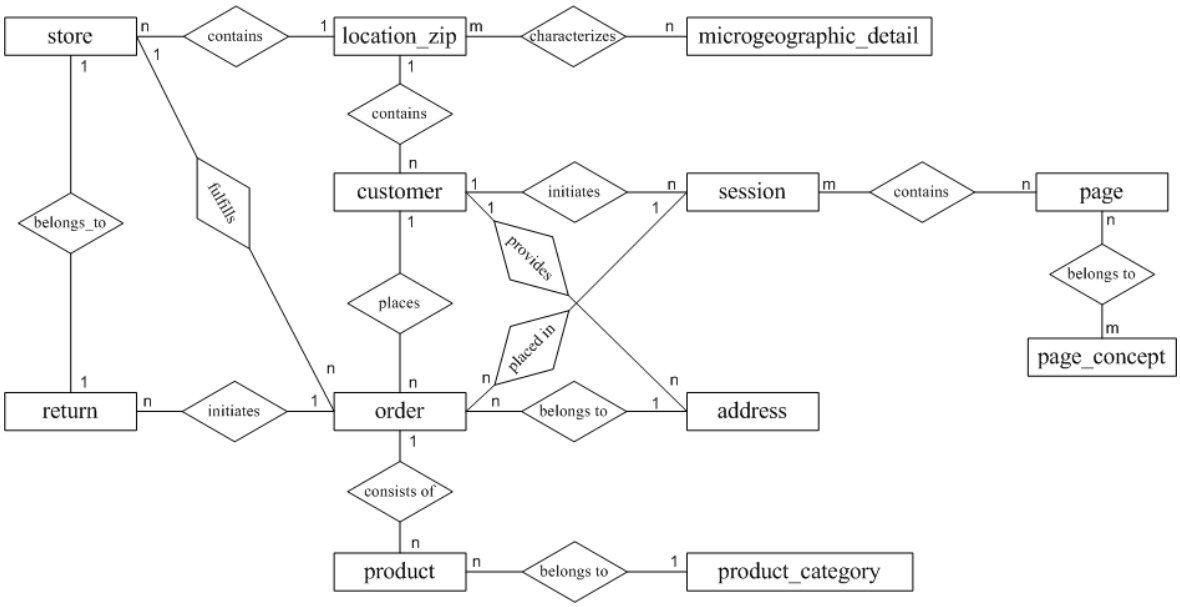

Figure 3-2: Entity relationship model of the multi-channel retailer

3.2 Framework categories

Related work has used the following terms for measuring notions of online success: “Web traffic measurements” [Malacinski, et al., 2001], “e-metrics” [Cutler and Sterne, 2000], “operational metrics” [Srivastava, et al., 2002], “metrics for Web merchandising” [Lee, et al., 2001], “visit related measures” [Moe and Fader, 2000], “CRM analytics” [SAP AG, 2001] and “Web log metrics” [Kohavi and Parekh, 2003]. Further terms of Web measurement have been introduced in Beal [2003], Bensberg [2001], Schwickert [2001] and Weigend [2003].

We use the following terminology in our framework: *Web metrics* are specific numbers or ratios assigned to a particular attribute (e.g. objects, events). Measurement techniques that cannot be expressed as a single number – e.g. distributions, association rules, or clusters – are referred to as *analytics*. The latter term is also used by many vendors of Web mining software [KDNuggets, 2005]. The term *Web analyses* is used as a superordinate label of both *Web metrics* and *Web analytics*.

Our analysis framework consists of five groups of Web analyses as depicted in Figure 3-3.

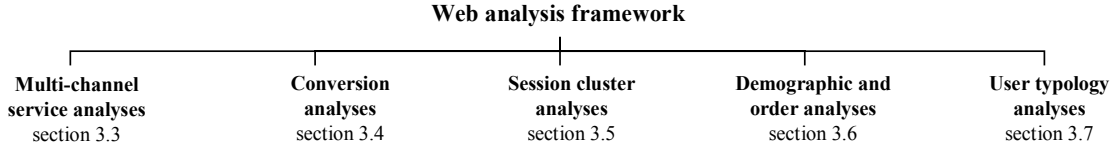

Figure 3-3: Framework categories

The five analysis categories address the notion of online success from different perspectives:

- First, Web sites must offer a flexible service mix in terms of convenient payment, delivery and return options to their customers in order to convince users to purchase online [Goersch, 2003; Omwando, 2002; USA Today, 2003]. Multi-channel Web sites may benefit from their ability to offer additional service options through a physical store network. We suggest a set of specific service analyses in order to measure consumer preferences of a Web site's multi-channel service offerings (cf. Section 3.3).
- Second, Web sites must increase the ratio of visitors to online buyers [Cutler and Sterne, 2000; Lee, et al., 2001]. This notion of success is also known as online conversion. On multi-channel Web sites, conversion does not measure a Web site's ability in attracting visitors to purchase at physical stores. Thus, we develop more fine-grained measures of conversion success in an Internet-only and multi-channel context (cf. Section 3.4).
- Third, a Web site must analyze the usage preferences of its visitors in order to improve site design and to derive information about a site's success in attracting specific groups of visitors [Moe, 2001]. We propose a session clustering approach that includes visitors' transaction and usage characteristics (cf. Section 3.5).
- Fourth, a Web site should focus its business efforts on the needs and preferences of those customers that are most valuable to the company. Thus, customer value needs to be defined. Indices measuring an online customer's value to a Web site are proposed. Moreover, customers are segmented according to demographic characteristics (cf. Section 3.6).
- Fifth, success needs to be evaluated in the context of non-commercial Web sites. An approach for measuring success incidents on information sites is proposed. The success of user search strategies on information Web sites will be analyzed (cf. Section 3.7).

The complete list of 82 metrics and analytics in the five analysis categories, their definitions, required data attributes and formalizations are depicted in Table 0-3 of the Appendix. All analyses are time-referenced. Sections 3.3-3.7 will present a selection of

these analyses and apply them on Web data from a multi-channel retailer and an information site.

Basic statistical aggregations of Web logs[4] (e.g. visits per day, distribution of user agents, most frequently visited Web pages, etc.) have not been integrated in our analysis framework as these analyses are offered by standard shareware tools [KDNuggets, 2005].

Moreover, product metrics and analytics are not presented in this thesis. Top-selling products and their position on a Web site are tracked routinely [Kohavi, 2004]. For example, market basket analysis is a common type of product data analysis that determines what products sell well together. A well-known algorithm for market basket analysis is the Apriori algorithm, which finds frequent itemsets in data [Agrawal, et al., 1993]. Linden et al. [2003] describe the recommendation algorithm used by the Internet retailer Amazon.com Inc.

Analyses calculating promotion or campaign success and cost-related analyses are also not included in the framework.

The proposed success analyses are particularly useful in the context of customer relationship management (CRM) [cf. Cutler and Sterne, 2000], Web site usability [cf. Kohavi and Parekh, 2003; Shneiderman and Plaisant, 2004; Spiliopoulou, et al., 2002a] and Web site personalization [cf. Kobsa, et al., 2001].

3.3 Multi-channel service analyses

This section presents metrics and analytics measuring consumers' service preferences for Internet-only and multi-channel retailers. Service offerings are considered one of the most important advantages of multi-channel over Internet-only retailers [Goersch, 2003; Omwando, 2002; USA Today, 2003]. A systematic analysis of service options in multi-channel retailing is presented in Section 3.3.1. The purchase decision process is introduced to point out multi-channel-specific service advantages. The current service mix of the 50 largest multi-channel retailers is presented in Section 3.3.2. The knowledge about the multi-channel service mix is used to define a set of service analyses in Section 3.3.3 and respective service metrics in Section 3.3.4. The metrics and analytics are applied on Web data from the multi-channel retailer. Section 3.3.5 concludes the

[4] E.g. unique visitors, page views, operating system, average time spent on pages, entry and exit pages, number of clicks or country code, search terms, referrers, server load, request errors, etc.

discussion of service preferences with a presentation of results from an online survey.

3.3.1 The multi-channel service mix

The purchase decision process is a well-known model that conceptualizes consumer choice as a number of predictable sequences of behavioral tasks in making purchases [Alba, et al., 1997; Engel, et al., 1968; Goersch, 2003; Howard and Sheth, 1969; Nicosia, 1966; Otto and Chung, 2000].

Figure 3-4 depicts an integrated view on the purchasing phases, which points out the main differences between Internet-only and multi-channel service offerings on Web sites.

Dotted arrows indicate the sales path at pure Internet retail sites. Continuous arrows indicate phase transitions at multi-channel retail sites where online customers can deviate from the Internet sales path and switch to traditional offline channels or back.

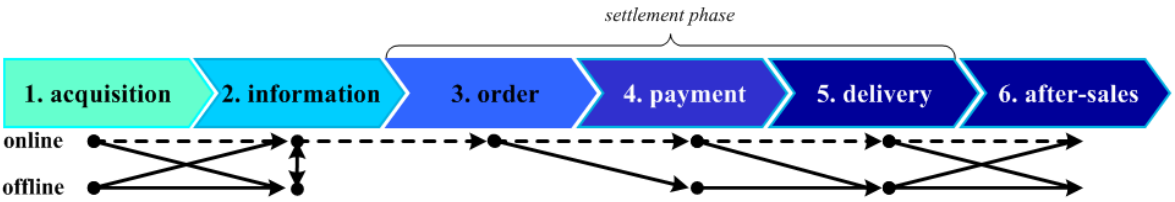

Figure 3-4: The purchase decision process at multi-channel and pure Internet retail sites

The names, number of tasks and labels of the purchasing process varies in the literature. The main difference between the models is their emphasis on different phases or the stress of specific cognitive aspects [Goersch, 2003]. Related models in an online context are the customer life cycle [Cutler and Sterne, 2000] and the customer buying process [Lee, et al., 2001], which will be discussed in more detail in Section 3.4.

The phases of the purchase decision process are used to systematically point out service advantages of multi-channel retail Web sites:

1. Acquisition (awareness) describes the phase where a consumer is attracted to a retailer's value proposition. In an online context, a click on the Web site would characterize the acquisition phase. An advantage of multi-channel Web sites is that consumers could be attracted to visit the Web site from physical stores (e.g. by using Internet terminals in stores).
2. During the information (persuasion) phase, visitors collect information about products and services and prepare their purchasing decision. In a multi-channel context, consumers could combine the advantages of online and offline information search. They can sample products in store after searching online, which may reduce the impediment of missing sensory clues on the Internet

[Rosen and Howard, 2000]. Moreover, multi-channel Web sites may support store-based search by displaying information about physical stores (e.g. opening hours, shop locations or product availability).

3. The first step of the settlement phase begins when a customer enters the order process. In an online context, the check-out of the shopping cart or input of customer data would characterize the commencement of the settlement phase.
4. In the payment phase, the customer initiates the payment of her order. Multi-channel retail sites can offer an additional payment option to their customers: customers may pay cash in-store after having ordered online.
5. Multi-channel Web sites can also offer more delivery options than pure Internet retailers. Online customers may pick up products in-store, which allows immediate gratification and avoids being present during the time of delivery. Some companies already offer special counters in stores where Internet orders can be picked up without waiting times.
6. During the after-sales phase, multi-channel retailers can provide an additional service to their customers: defect or unsatisfactory orders may be returned in physical stores, which could be more convenient than returns by mail. Multi-channel Web sites may also offer additional assistance (e.g. maintenance, installations) executed by personnel from nearby physical stores.

3.3.2 Site services in multi-channel retailing

The analysis of multi-channel characteristics in the customer purchasing process facilitates the identification of five additional service options that can be offered on multi-channel retail sites:

- *in-store payment*: online orders can be paid in a physical store.
- *in-store pickup*: visitors may place an order online, but pick up products in a physical store.
- *in-store returns*: online orders can be returned in a physical store
- *store locator*: multi-channel retailers can offer pages where online visitors can find information about physical stores (e.g. opening times, addresses, maps) in their neighborhood.
- *inventory check*: site visitors may check inventory or search for special offers in stores.

We observed the availability of these service options at the world's 50 largest e-retailers in 2002 [Gallo and McAlister, 2003]. 43 of these e-retailers operate multiple distribution channels, seven are pure Internet-players. From the 43 multi-channel retailers, 30 operate

physical stores[5] and 13 primarily operate direct distribution channels such as catalogs, TV or call centers. In Table 3-3 we give an overview of the present service mix at the 30 retailers that operate physical stores and a Web site:

number of retailers	in-store payment	in-store pickup	in-store return	store locator	inventory check
3	▮	▮	▮	▮	▮
2	▮	▮	▮	▮	▮
4			▮	▮	▮
10			▮	▮	
2				▮	▮
9				▮	

Table 3-3: Online service mix at the 30 largest multi-channel retailers (as of November 2003)

The analysis indicates that many retailers do not offer the full multi-channel service spectrum. The most common service combination includes store locator pages and in-store returns of online orders. All multi-channel retailers in the sample offer store locator pages and about two-thirds offer in-store returns[6]. At eleven companies online customers can check store inventory and/or special offers in physical stores. At five companies customers can pick up online orders in physical stores. Three companies offer the full multi-channel service spectrum including payment in-store after an order has been placed online.

Whereas returning goods from online purchases back to a physical store is a typical service option at many multi-channel retailers, the practice of picking up goods or checking stock in a particular store is less common, yet.

A retailer's choice of a particular service mix may depend on several parameters. A large store network seems to be a requirement for in-store pick-ups. Retailers offering the full multi-channel service spectrum operate a nationwide retail network. Moreover, differences between online and offline pricing present a challenge to multi-channel integration. Local

[5] Only those retailers with a large number of stores were counted as retailers operating physical stores.

[6] A recent study found that 78 percent of retailers offer in-store returns of online purchases (Shop.org 6.0).

discounts at stores could confuse online customers when they pick up orders and recognize a lower in-store price. A study found that one-third of multi-channel retailers offered different online and offline prices in 2001 [Shern, 2001]. Some of the multi-channel retailers in our sample announced on their Web site that any discounts in-store also apply to online orders on the day of pickup (e.g. Circuit City Inc.). Delivery cost is a further decision parameter that needs to be considered in multi-channel retailing. The avoidance of shipping cost is one of the most important reasons for online customers to pick up orders (cf. Section 3.3.5). Customers of online retailers offering low shipping costs or free-of-charge delivery may have fewer incentives to use an in-store pickup service. Cost for order management and additional personnel could be a further reason why many multi-channel retailers have not yet fully integrated online and offline services.

As this brief discussion has demonstrated, a retailer's decision to offer multi-channel services is influenced by many organizational parameters. An in-depth discussion of these parameters is not within the scope of this work.

3.3.3 Service analytics

Our analysis of the service mix constitutes the basis for the definition of a set of service analyses measuring consumer service preferences in multi-channel retailing. The analyses are applied on data from the multi-channel retailer, who offers an integrated service spectrum in the sense of Table 3-3 except that a search function for in-store inventory is not yet implemented on the Web site. Online customers can pay online by credit card, directly at a physical store or by cash on delivery. Online orders are delivered directly to the customer or can be picked up at a store. Returns can be handled either by mail or at a physical store. Visitors can locate the nearest store online.

We analyzed data from 13,653 customers who made 14,957 transactions over a period of 8 months in 2001/02.

The service analytics are presented as association rules, which depict relationships among items based on their patterns of co-occurrence across transactions [Agrawal, et al., 1993]:

Let $I = \{I_1,\ldots,I_n\}$ be a set of discrete entities (items) and $D = \{t_1,\ldots,t_k\}$ a set of transactions in a database D with $t \subseteq I$. Then $X \Rightarrow Y$ is a association rule with $X \subseteq I$, $Y \subseteq I$, $X \cup Y = \varnothing$.

$X \Rightarrow Y$ has support s if s% of transactions in D contain $X \cup Y$.

$$support\,(X \Rightarrow Y) = \frac{\left|\{t \in D \middle| X \cup Y \subseteq t\}\right|}{\left|D\right|}$$

The rule confidence *c* is defined as:

$$confidence\ (X \Rightarrow Y) = \frac{support\ (X \Rightarrow Y)}{support\ (X)}$$

The presentation of service preferences as association rules provides two benefits: first, the Web analyst can easily identify the most important service rules and second, the frequency of occurrences between service offerings can be depicted concisely.

3.3.3.1 Payment and delivery preferences

The first set of association rules describes the associations between customers' payment and delivery preferences (cf. step 4 and 5 of the customer purchasing process in Figure 3-4):

(1) *Online payment* $\Rightarrow_{s=0.27,\ c=0.97}$ *Direct delivery*

(2) *Online payment* $\Rightarrow_{s=0.02,\ c=0.03}$ *In-store pickup*

(3) *Cash on delivery* $\Rightarrow_{s=0.02,\ c=0.06}$ *Direct delivery*

(4) *In-store payment* $\Rightarrow_{s=0.69,\ c=0.94}$ *In-store pickup*

The first row would be read as follows: if a customer chose online payment using a credit card, she also chose direct delivery with 97% frequency. This rule could be identified in 27% of the transactions. Thus, 3,686 orders were delivered directly. Surprisingly, in 69% of the transactions, customers placed an order online but chose to pay and pick up their order at a physical store (rule 4). Several surveys confirm that observation even though with a lower support factor [Swerdlow, et al., 2002; Tedeschi, 2001]. In 27% of the transactions, customers chose the service combination of online payment and direct delivery – the typical service combination offered at pure Internet retailers. Only very few customers tend to combine online payment and in-store pickup (rule 2). Moreover, only a few customers paid cash on delivery (rule 3). For comparisons, in Germany, 64% of e-commerce offers are purchased on account, 36% by payment on delivery, 26% by direct debit and 23% by credit card [Schneemann, 2003].

As a conclusion, most online customers collect information and place orders on the multi-channel site but prefer physical stores for pickup and payment. Less than one-third of the customers in the sample are "pure" online users who chose direct delivery and online payment.

3.3.3.2 Return preferences

Moreover, we analyzed customers' return preferences at the multi-channel retailer. 10% of all online orders were returned within eight months. The association rules (5) and (6) represent customers' return preferences (cf. step 5 and 6 of the customer purchase process presented in Figure 3-4):

(5) *Return* $\Rightarrow_{s=0.06,\ c=0.87}$ *In-store*

(6) *Return* $\Rightarrow_{s=0.04,\ c=0.13}$ *Mail-in*

The findings indicate a strong preference for in-store returns (87%). Though returns were offered free of charge, only 13% of all returned orders were mailed back. The customers who returned orders by mail also had chosen online payment and direct delivery when they placed their order. A consumer survey found similar results: 83% percent of online buyers would prefer to return online purchases at stores [Jupiter Research Corporation, 2001].

A reason for the preference of in-store returns could be the convenience of personal assistance and the handling of packaging in-store. Moreover, replacement or guarantee issues can be discussed in person in-store. The offer to return online orders at a physical store seems to be a successful service strategy that is offered by two-thirds of the largest multi-channel retailers (cf. Section 3.3.2).

3.3.3.3 Repeat customers' service preferences

The last set of association rules describes the migration behavior of repeat customers' delivery and payment preferences. Migration measures the number of customers who switched their delivery or payment preferences in at least one transaction after their first one. The number of repeat customers amounts to 10% of all customers over a time period of eight months. Only 9% of repeat customers changed delivery terms after their first transaction. None of the customers switched their transaction preferences more than once.

(7) *Direct delivery* $\Rightarrow_{\mathbf{s=0.001,\ c=0.15}}$ *In-store pickup (in ≥1 of the following transactions)*

(8) *Direct delivery* $\Rightarrow_{\mathbf{s=0.003,\ c=0.85}}$ *Direct delivery (in every following transaction)*

(9) *In-store pickup* $\Rightarrow_{\mathbf{s=0.001,\ c=0.10}}$ *Direct delivery (in ≥1 of the following transactions)*

(10) *In-store pickup* $\Rightarrow_{\mathbf{s=0.004,\ c=0.90}}$ *In-store pickup (in every following transaction)*

The support for repeat customers who switched to in-store pickup (rule 7) was equal to

the support for customers who switched to direct delivery (rule 9) in at least one of the following transactions after the first one.

As payment and delivery preferences are closely coupled (cf. rules (1)-(4)), the support and confidence values for payment migration between online payment and payment in-store were equivalent to rules (7)-(10) in our sample.

Rule (9) could be interpreted as an indicator of trust in the online shop: if an online customer picks up or pays a product in-store first and then switches to direct delivery or online payment, the consumer may have developed trust in the retailer's direct delivery and online payment reliability.

3.3.4 Service metrics

The service rules of Section 3.3.3 can be transformed into service metrics that are simple to calculate and can be easily used for comparisons over time and between Web sites. Table 3-4 presents a list of multi-channel-specific service metrics and their results that can be derived from the association rules presented in Section 3.3.3.

Multi-Channel Service Metrics	Results
In-store payment rate	= 69%
Online payment rate	= 29%
Cash-on-delivery payment rate	= 2%
In-store payment migration rate	= 15%
Online payment migration rate	= 10%
Deliveries-to-stores rate	= 71%
In-store delivery migration rate	= 15%
Direct delivery migration rate	= 10%
Returns-to-stores rate	= 87%

Table 3-4: Multi-channel service metrics

The in-store payment rate measures the number of online customers who paid in-store and is equivalent to the support factor of association rule (4). The online payment rate measures the number of online payers and is equivalent to the sum of the support factors of rules (1) and (2). The cash-on-delivery payment rate is the support factor of rule (3). The deliveries-to-stores rate measures how many customers preferred to pick up their

online orders at physical stores. It is the sum of the support factors of association rules (2) and (4). The returns-to-stores rate measures how many buyers returned products in physical stores. It is the confidence factor of rule (5).

The in-store delivery migration rate measures the number of repeat customers who switched from direct delivery to pickup in-store in at least one of their following transactions. It is equal to the confidence factor of association rule (7). The result of the direct delivery migration rate is equivalent to the confidence factor of rule (9). The payment migration rates are calculated analogous to the delivery migration rates.

3.3.5 Survey results

To round up our analysis of service preferences we conducted an online survey on the multi-channel Web site to inquire reasons for the surprisingly high number of in-store pickups. Consumer comments from a previous survey[7] were consulted to define seven answer options to the question “if you have decided to pick up an online order at the retailer, what were the reasons?”. This question was attached to the online questionnaire described in Chapter 2. 1048 visitors checked 3505 answer fields. The results are depicted in Figure 0-2 of the Appendix.

The survey results show that shipping costs are most important for customers to pick up orders. The retailer’s shipping cost are 4.95 euros and thus below the German average for domestic postal ground shipping of consumer electronics. Costs are waived for orders equal to or greater than 100 euros. The retailer offers standard delivery times of about three days.

The second most important reason to pick up orders in physical stores was the need to look at the product in person and the demand of direct communication. Half of the users prefer to look at a product before they accept it and 41% want to see that a product is not damaged. Delivery convenience and online payment risks are also significant reasons to pick up orders in-store. 26% claim they are usually not at home during delivery times and 20% pick up orders to avoid the lag time of shipping. 19% find online payment too risky.

3.3.6 Summary and implications

The multi-channel service mix at the top 30 multi-channel e-retailers in 2002 has been

[7] The survey was placed on the Web site in 2002 [Teltzrow and Berendt, 2003]. 4267 respondents gave open text answer to the question “what do you like/dislike about this Web site”. 345 answers addressed multi-channel services.

analyzed. The results demonstrate that Web sites increasingly extend multi-channel services to their customers. In particular, in-store returns and a store locator are typical service options at large multi-channel retailers. However, the analysis has shown that many companies do not yet fully exploit the potential of multi-channel service integration. The analysis of consumer preferences demonstrated a clear demand for such services, however.

In order to measure these service preferences, a group of service analyses has been presented. The results indicate that consumers have a strong preference for in-store pickup, payment and return.

The presented service analytics and metrics have important implications for business decision making. For example, if a large percentage of users prefers to examine and to pick up products in store, it may be worthwhile to further expand the store network.

3.4 Conversion analyses

This section focuses on the analysis of Web usage behavior and presents a set of Web analyses measuring fine-grained conversion metrics for Internet-only and multi-channel retailers.

Conversion – defined as the proportion of visits that end with a purchase – is a well-known notion of online success. The online conversion rate for US retailers increased from 2.2% in 2000 to 3.1% in 2001 [BCG and Shop.Org, 2002]. However, only 2–3% of user sessions are captured in this success metric, whereas 97-98% of session data stem from visitors who looked at information on the Web site but did not engage in an online transaction. The session data from this latter group may provide useful insights in alternative success incidents on Web sites though. Moreover, a single conversion rate is not sufficient for measuring the success of multi-channel Web sites: in a multi-channel context, conversion success may not be visible directly in the Web logs, e.g., if visitors collect information online but purchase offline. Thus, more fine-grained conversion metrics need to be developed.

In Section 3.4.1, we introduce the customer life cycle of Cutler and Sterne [2000] and the micro-conversion rates of Lee et al. [2001] and derive a formal model measuring conversion success in Internet retailing. We will refer to techniques from Web *usage* mining, which is the application of data mining techniques to discover interesting Web usage patterns [Baldi, et al., 2003; Cooley, et al., 1999; Han and Kamber, 2000; Kosala and Blockeel, 2000; Spiliopoulou and Faulstich, 1998; Srivastava, et al., 2000].

Section 3.4.3 presents new conversion success metrics: a class of *concept conversion*

rates, and the *offline conversion rate*, that provide a fine-grained view on consumers' conversion behavior. In order to calculate these metrics, a taxonomy of site concepts for the multi-channel retailer has been developed.

In Section 3.4.4, we calculate the conversion metrics and discuss our results. Recommendations for site improvement are derived.

3.4.1 Conversion success metrics

The processes whereby a visitor becomes a customer (cf. Section 3.3.1) have been described for an online retail context in related work: on a macro level, the processes of moving along the customer life cycle [Cutler and Sterne, 2000]; on a micro level, the processes of moving along the customer buying process [Lee, et al., 2001]. In each of these models, distinct stages (and user groups who are defined by having "reached" those stages) follow upon one another. In [Berthon, et al., 1996], the purchase process is modeled by distinguishing, within the set of all site users, the "short-term visitors" from the "active investigators". Some of the latter eventually become "customers". Metrics are proposed to measure how many site users reach these advanced stages. However, to find out why short-term visitors may not have become active investigators, or active investigators may not have become customers, it is necessary to consider the visited pages with respect to their potentials for further action. Criteria for classifying pages accordingly can be based on merchandizing purpose [Lee, et al., 2001] or, more generally, on service-based concept hierarchies [Spiliopoulou and Pohle, 2001]. The paths taken to goal pages, their lengths in particular, have been integrated as a further aspect of efficiency control [Spiliopoulou and Berendt, 2001].

3.4.2 An integrated framework for conversion success

As a first step towards a model of conversion success measurement, an integrated scheme for formalizing both the life-cycle metrics of Cutler and Sterne [2000] and the micro-conversion rates of Lee et al. [2001] has been proposed. Figure 3-5 illustrates the stages and processes of these models. The figure should be read as follows: the letter at a node identifies a set of people defined with reference to the site's goal.[8] The subscript *T*

[8] For simplicity, we assume that in a given time period T, there is only one goal, or several which can be aggregated into one goal along concept hierarchies. This framework treats all users who visit a given class of pages as equal. It may be argued that this represents a simplified description of the complex goal-setting and decision-making processes that users go through when navigating a site. However, this simplification is justified by the purposes of a business-related outcome analysis.

is omitted in the figure to enhance clarity. By the actions performed in *T*, each individual moves from being an element of the set at one node to being an element of either of the sets at the children of that node. For example, all "suspects" $\in S_T$ (i.e., people who have become aware of the site and are visiting it [Berthon, et al., 1996] are either "acquired" and become "prospects" $\in P_T$ (i.e., people who show interest by some kind of active participation, cf. the "active investigators" of Berthon et al. [1996] or not. In the latter case, they belong to the set *nPT*. Children of a node partition the set of their parent node, e.g., $P_T \cap nP_T = \emptyset$, $P_T \cup nP_T = S_T$. Figure 3-5 (a) shows the stages and transitions involved in the life-cycle metrics of Cutler and Sterne [2000], and (b) shows an alternative partitioning of the set *CT* of customers in (a). That is, it is possible that $U1_T \cap C1_T \neq \emptyset$, $U1_T \cap CA_T \neq \emptyset$, $U1_T \cap CR_T \neq \emptyset$ und $UR_T \cap C1_T \neq \emptyset$, $UR_T \cap CA_T \neq \emptyset$, $UR_T \cap CR_T \neq \emptyset$. Figure 3-5 (c) shows a more fine-grained representation of the stages of the customer buying cycle that make up the steps that convert a prospect into a customer.[9]

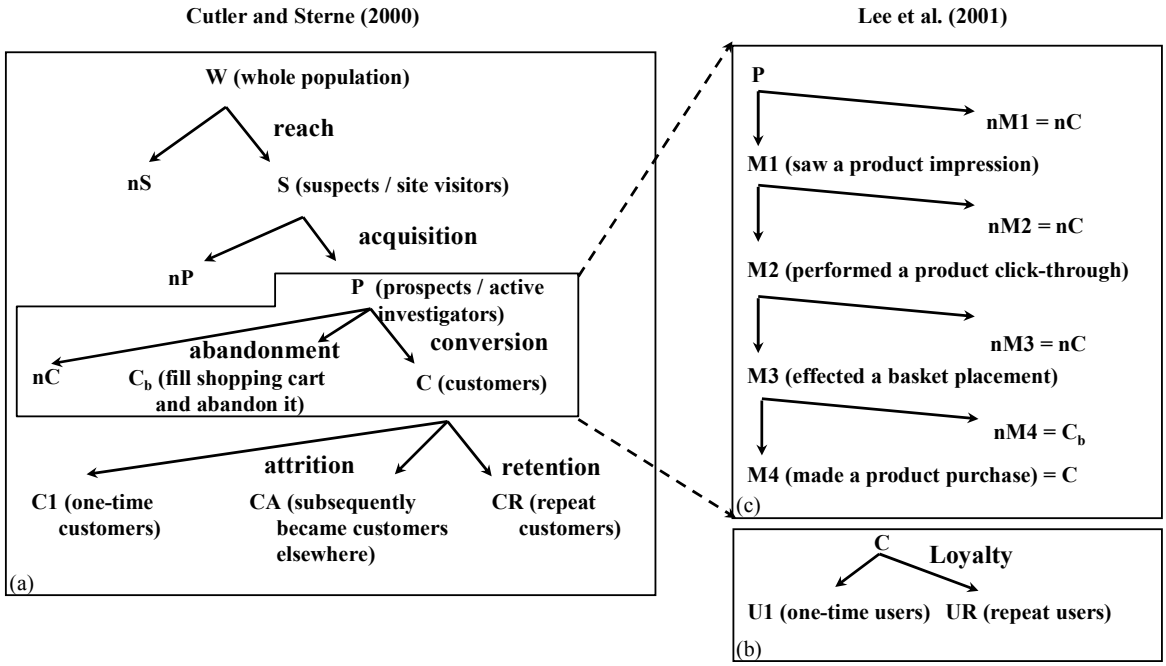

Figure 3-5: (a), (b): Stages and transitions in the customer life cycle, and (c) in the customer buying cycle

In Table 3-5, we propose formalizations of the metrics associated with the transitions of

[9] Note that conversion, abandonment, etc. are defined relative to the site's goal, so "customer" in the general case means "person who reached the site's goal", and "abandonment" means "abandoning a task on the site whose completion constitutes the site's goal".

the customer life cycle [Cutler and Sterne, 2000] in Figure 3-5 (a) and (b), and we express the micro-conversion rates of Lee et al. [2001] (Figure 3-5 (c)) in the same framework. This representation assumes that to become a customer, one must follow the canonical sequence shown in Figure 3-5 (c).

The last column of Table 3-5 points out data requirements for calculating the metrics. If the rate of visits that lead to active participation is of more interest than numbers of individual customers, session IDs suffice, and acquisition can be measured as the number of visits with URL requests that indicate active participation, divided by the number of all visits, in *T*. Conversion and abandonment can be measured analogously, cf. Spiliopoulou and Berendt (2001) and Spiliopoulou and Pohle (2001) for examples. Measures like retention or attrition, of course, rely on the personal identity of the customer and therefore require at least cookie data as (quasi-)unique customer identifiers. *Reach* requires marketing data about the number of Internet users and the overall size of the target market.

Life Cycle Metrics	Metrics Definition	Data Requirements
Reach	S_T / W_T	M
Acquisition	P_T / S_T	C (SI)
Conversion	C_T / P_T	C (SI)
Retention	CR_T / C_T	C and/or TA
Loyalty	UR_T / C_T	C
Abandonment	Cb_T / P_T	C (SI)
Attrition	CA_T / C_T	TA, M
Churn	$\frac{\lvert CA_T \rvert}{\sum_{t=1}^{T} (\lvert C_t \rvert - \lvert CA_t \rvert)}$	TA, M

Micro-Conversion Rates		
Look-to-click	$M2_T / M1_T$	C (SI)
Click-to-basket	$M3_T / M2_T$	C (SI)
Basket-to-buy	$M4_T / M3_T$	C (SI)
Look-to-buy	$M4_T / M1_T$	C (SI)

M = marketing, C = cookies, SI = session ids, TA = transaction

Table 3-5: Metrics for e-business: life-cycle metrics and micro-conversion rates

3.4.3 New conversion metrics

The formalization of the micro-conversion rates of Lee et al. [2001] presents two problems:

Problem 1. Although these metrics are useful for determining specific site events, the four conversion rates proposed by Lee et al. do not look more detailed into the users' information behavior such as a user's clickstream from a catalog site to a product page. In particular, they lack a definition of conversion in the context of multi-channel retailing.

Problem 2. The proposed conversions have been defined on the basis of sessions that reach the next phase in the buying process or not. However, they do not consider volume-based conversion (how many pages representing one phase have been visited relative to those representing another phase).

Our approach addresses these two issues. First, we use an OLAP-style analysis to address problem (1). We suggest a general formalization of fine-grained conversion rates that can be used on different Web sites. We develop and use a concept hierarchy to achieve a more aggregate view of the data, and we extend the classification of pages by merchandizing purpose to also measure cross-channel affinity. We then investigate session modeling in order to address problem (2), using feature vectors that indicate either whether a concept has been visited in a session or not, or how many times it has been visited. We use sessions instead of users as our basic unit of analysis because our focus is on the micro level of individual online interaction processes, rather than on the macro level of how a person moves along the customer life cycle. Session-based analysis has been shown to be useful for a number of applications such as recommender [Sarwar, et al., 2000] and personalization [Kobsa, et al., 2001; Mobasher, et al., 2002] systems. Moreover, session-based data collection (or the reconstruction of sessions from IP+agent) presents fewer privacy problems than cookie-based data collection, which will be

discussed in more detail in Chapter 5. Furthermore, cookies can be deleted, which impedes a re-identification of users [Fiutak, 2004]. However, the use of session IDs assumes that each session originated from a different user, which must not be true.

3.4.3.1 Multi-channel site taxonomy

For incorporating domain knowledge in the log analysis, we built a concept hierarchy as a model of the business purpose underlying the multi-channel Web site introduced in Section 2.3.1.

A concept hierarchy, also known as *taxonomy*, generalizes concrete objects into more abstract concepts [Berendt and Spiliopoulou, 2000; Pohle and Spiliopoulou, 2002; Spiliopoulou, 2000]. The development of concept hierarchies requires the mapping of user activities into generic user tasks. This procedure provides two main benefits: first, previous knowledge about a site's business objectives can be integrated in the analysis process. Second, the data are much easier to interpret by the analyst, e.g. statistical analysis can be performed on product group rather than product level.

The mapping of site components to concepts is traditionally performed prior to the statistical analysis of the data. The establishment of a concept hierarchy cannot be automated, since the site semantics depend on the goals of the Web site and the objectives of the institution owning it. E-commerce sites usually have well-structured Web content, including predefined metadata or a database schema [Lynch and Horton, 2001; Shneiderman, 2000; van Duyne, et al., 2002].

Our classification covers the types of services that typically constitute a large multi-channel retail site. It extends the usual classification of the purchase decision process (cf. Figure 3-4) by a more fine-grained concept view that includes the *service*, *offline information*, *information catalog* and *information product* concept. The following concepts are included in the taxonomy:

1. *acquisition (home)*: all Web pages that are semantically related to the initial acquisition of a visitor (e.g., the home page).
2. *information catalog (infcat)*: pages providing an overview of product categories. This concept could be further differentiated with a number of sibling nodes describing the Web retailer's product categories.
3. *information product (infprod)*: pages displaying information about a specific product. *infprod* is a child of *infcat*.

4. *service*: general company information, registration, games and other trust-building information.
5. *transaction*: all transaction pages before an actual purchase, starting with a customer entering the order process, check-out of shopping cart, input of customer data, payment and delivery preferences.
6. *purchase*: pages indicating the completion of the transaction process such as the invocation of an order confirmation page.
7. *offline*: all pages related to any offline information: store locator (pages for finding physical stores in one's neighborhood), information about offline services, or specific offline referrers.[10]

Figure 3-6 depicts the site taxonomy that was used for the analysis.[11] Each of the 760,535 page requests that remained after data preprocessing were mapped onto concepts from the hierarchy.

Based on this categorization of pages, we propose *concept conversion rates* as ratios of page impressions between two concepts. Ideally, high transition rates between adjacent phases should be achieved.

[10] Offline referrers are visits from referring URLs that are uniquely linked to offline stores, such as hits from affiliated stores that provide specific URLs to the main Web site.

[11] More fine-grained taxonomies have been developed. However, the depicted aggregation suffices our analyses purposes.

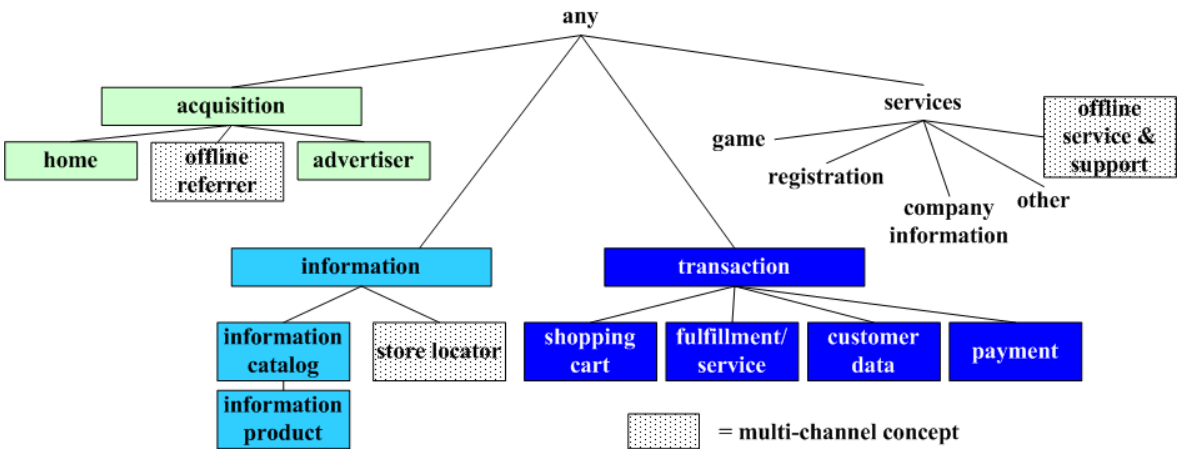

Figure 3-6: Site taxonomy

3.4.3.2 Conversion rates and visit rates

Sessionized data can be analyzed in a number of ways. A session is usually treated as a *bag* of visited pages or visited page concepts, as a *set*, or as a *sequence*. Here, we will focus on analyses of bags or sets, which are useful for applications like market basket analysis and recommendation systems based on analyzing pages that were accessed together in users' previous sessions [Cutler and Sterne, 2000; Perkowitz and Etzioni, 1998; Zaiane, et al., 1998]. Each session s from S, the set of all sessions, can then be represented as a feature vector (cf. Section 3.1.1 for a formal definition) with each component $s[c]$, c=1,…,7 indicating either the number of visits to the respective concept 1–7 (bag), or, in a dichotomized fashion, stating whether or not that concept was visited in the session (set). In the following, we will refer to the first method as *weighted-concept* and to the second as *dichotomized-concept*, with $s_w[c] \in N_0$ and $s_d[c] \in \{0,1\}$. In addition to concepts 1.–7., $s_d[0]$ denotes the visit to "any" concept, i.e., $\{s \in S | s_d[0] = 1\} \equiv S$.

We first define the *dichotomized-concept conversion rate* from concept c_i to concept c_j as

$$c_i\_to\_c_j^d = \frac{\{s \in S | s_d[c_j] = 1 \,\&\, s_d[c_i] = 1\}}{\{s \in S | s_d[c_i] = 1\}}$$

$$= \frac{\sum_{s \in S} And(s_d[c_j], s_d[c_i])}{\sum_{s \in S} s_d[c_j]}. \qquad (1)$$

This notation shows that the conversion rate can also be read as the confidence of the *association rule* $c_i \rightarrow c_j$.

Two cases can be distinguished. The first assumes that a visit to concept c_j is only possible *after* a visit to concept c_i. In this case, equation (1) can be simplified. Abbreviate the denominator as S_i, and define S_j, $S_{i\&j}$ analogously. Then, because $S_j \subseteq S_i$,

$$c_i \_to\_ c_j^d = \frac{|S_{i\&j}|}{|S_i|} = \frac{S_j}{S_i}.$$

Examples are the conversion rates shown in Table 3-5. In this fashion, one can also address the question whether a visit accessed a particular concept c_i at all. This gives rise to *total conversion rates* c_0 to c_i, which means that the denominator becomes $|S_0| \equiv |S|$. We specify this for the *offline* concept. Let $S_{offline} \in S$ be the set of sessions that visit the *offline* concept at least once, i.e., $S_{offline}=\{s \in S | s_d[offline]=1\}$. Then we define the *offline conversion rate* as $(|S_{offline}|/|S|)$. We add a second case, which concerns two concepts that need not necessarily occur in the order *i*, *j*. An example is the prodinf_to_service conversion rate that we will investigate in the next section. Furthermore, we extend this analysis by a set of volume-based metrics. We define the *weighted-concept visit rate* from concept c_i to concept c_j as

$$c_i \_to\_ c_j^w = \frac{\sum_{s \in S} s_w[c_j]}{\sum_{s \in S} s_w[c_i]}. \qquad (2)$$

While this cannot directly be broken down to the number of concept visits occurring within the same sessions (and thus does not describe the conversion of one visitor from being in one subgroup of *S* to being in another subgroup), it is a useful indicator of the different concepts' relative importance throughout the whole log. The idea of using a concept hierarchy for analysis can be extended by further partitioning these sets. For example, we investigated the set of *sessions that visit the store locator, SLV,* and the set of *sessions that exit via the store locator, SLE.* Both are dichotomized-concept notions, and $SLE \subseteq SLV \subseteq S_{offline}$. Finer-grained offline conversion rates can be calculated using these sets.

Visits to concepts and conversion rates not only produce numbers for eventual success measurement. They can also be used to gain insights into online users' behavior, in particular if different groups of users are compared. In Section 3.4.4, we illustrate how the computation of concept visit frequencies and conversion rates can help to understand the use of a multi-channel Web site not only within the set of all sessions *S* as in equations (1) and (2), but also in other base sets.

3.4.4 Conversion metrics results

We modeled the visits in terms of the concepts introduced in Section 3.4.3.1 and computed the conversion rates defined in Section 3.4.3.2.

We first compared two groups of sessions: the set of all sessions *S* and the set of all purchase sessions *C*. Moreover, we differentiate between two multi-channel-specific session groups: within the set of purchase sessions, we compare the two groups with the different delivery choices *pick-up in store* and *direct delivery*. We use delivery choice as an exemplary feature of multi-channel affinity because Section 3.3 has shown that delivery services are one of the most important service advantages of multi-channel retailers over pure Internet merchants. The purchase behavior of these groups is particularly interesting as one group uses the direct delivery option preferred by traditional Internet shoppers whereas the other demonstrates a multi-channel affinity.

The first group is obtained from the Web logs, and the other three groups are obtained by (a) combining Web log data with the transaction back-end data, and (b) classification according to the values of the relevant attributes (purchase: yes/no, delivery choice: direct delivery/pick-up in store).

Figure 3-7 (a) shows the numbers of page impressions on the various concepts in the set of all sessions *S* and Figure 3-7 (b) the set of all purchase sessions.

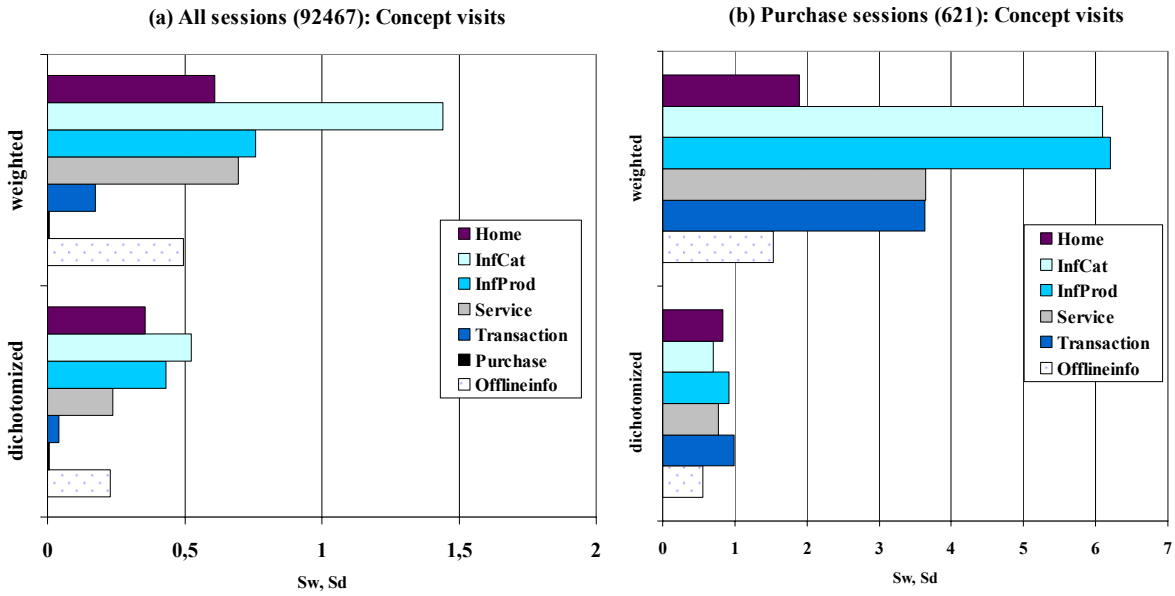

Figure 3-7: (a) all sessions and (b) purchase sessions: normalized numbers of weighted and dichotomized concept visits per session

The upper bars show the average number of visits, in one session, to each of the 7 concepts, and the lower bars show the proportion of sessions that have visited each of the 7 concepts at least once. For example, the *infcat* concept was visited, on average, 1.44 times per session, but in fact only 52.5% of all sessions visited this concept at all. Visit rates correspond to the relative widths of the "weighted" bars. This normalization was done to allow the best possible comparison between usage behavior in the four groups of sessions we investigated (cf. Figure 3-7 and Figure 3-8).

The findings from this analysis suggest that not all sessions include the *home concept*. Some visitors follow links from affiliate sites that often lead directly to the *infprod* concept. As expected, most hits occur in the information phase, where users explore product information before they eventually visit service-related sites, purchase a product or leave the site. One-fourth of all user sessions visited the offline concept at least once. The conversion rates are based on single-session conversion from one concept to another, but they lack the volume information. Especially in a multi-channel context, the information on volume combined with the offline conversion could indicate that the site serves information needs and increases the interest in offline sales. Low visit rates indicate that one should look at data on a more detailed level to identify inefficiencies within certain site concepts. Figure 3-7 (b) shows the normalized numbers of page impressions on the various concepts in the set of all purchase sessions *C*. The *purchase* concept is not shown because it is, by definition, always visited.

The comparison with the group of all sessions indicates that users who decide to initiate a

purchase do this on a basis of a much more extensive interaction with the site. In particular, the total number of catalog and product information pages visited are much higher, on average, in a purchase session. Not surprisingly, nearly every purchase was preceded by a visit to a product information page. Service was used more often in purchase sessions. Offline pages were also visited by more than 50% of the user sessions.

Figure 3-8 shows the purchase sessions with direct delivery and pick-up preferences. The results are based on a sample of 621 transaction records that have been linked to the respective Web-usage records. Session IDs were used to link the purchase sessions and transaction records (cf. Section 3.1). 326 users preferred direct delivery, whereas 295 preferred pick-up in store.

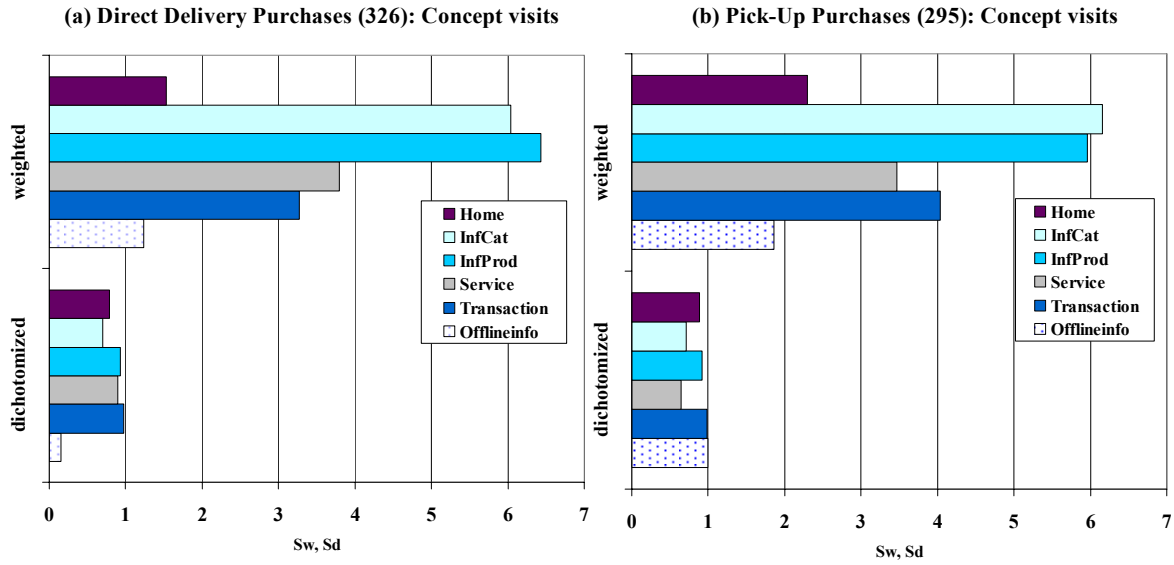

Figure 3-8: (a) Direct delivery purchase sessions and (b) pick up purchase sessions: normalized numbers of weighted and dichotomized concept visits per session

The 326 sessions with direct delivery preference differed in their navigation behavior from the 295 sessions with pick-up in store preference. Figure 3-8 (a) and Figure 3-8 (b) illustrate the two subgroups' concept visits. The figures show that the behavior is generally very similar, in particular when one looks at the dichotomized concepts. However, there are two key differences. Nearly all people with pick-up preference looked at offline concepts: they located the nearest shop. In contrast, for customers who chose direct-delivery, the service concept was very important; most probably serving a trust-building function.

The concept conversion rates summarizing this comparison between all four session groups are shown in Table 3-6.

Base set	H→IC	IC→IP	IP→TA	TA→S	OCR
all	0.75	0.5	0.06	0.23	0.23
purchase	0.8	0.95	0.98	0.77	0.56
direct delivery	0.82	0.96	0.97	0.89	0.16
store pick-up	0.78	0.93	0.99	0.64	0.997

H = home, IC = infcat, IP = infprod, TA = transaction, S = service, OCR = offline conversion rate

Table 3-6: Selected conversion rates in the four sets of sessions

We also investigated in more detail the store locator visits. We found that in the set of all sessions 13% of all user sessions included at least one invocation of the store locator concept (SLV=13%). This number demonstrates the importance of the multi-channel concept. For more than 6% of the sessions, pages belonging to the store locator were used as the exit page (SLE=6%). This indicates a group of visitors that collects information online before locating the next store. The store locator was also the concept with a high percentage of one-click visitors (12.5%). The behavior pattern of one-click visitors on the shop locator is interesting as it indicates shoppers who are solely interested in finding the next retail store. Thus, they use the Web as a type of "yellow pages".

3.4.5 Summary and implications

In the Web, unlike in a physical store, it is feasible and economical to measure conversion at a much finer level of detail; the inspection of path-dependent conversion rates may therefore yield valuable insights into a retailer's success in funneling consumers through a Web site before a purchase is made. From a marketing point of view, the proposed metrics provide site managers with arguments why a Web site contributes significantly to a retailers overall success even though this might not be reflected in actual Web sales figures. Fine-grained conversion rates allow the analyst to determine bottlenecks in the buying process and the newly introduced offline conversion rate is an indicator for the site's success in inducing offline sales.[12] The overview of Web metrics, their requirements and potential uses provides site analysts with a platform to efficiently determine

[12] It could be supplemented by retailers who track the number of visitors who come into a physical shop with a printout from the Web site.

conversion success.

In the case of the multi-channel retailer, the results indicate that (a) purchase sessions have a much “broader funnel” than the average session, i.e., the large majority of users in purchase sessions proceed from each step to the subsequent one. (b) For sites with high percentages of direct delivery preferences, it is very important to maintain helpful service pages. (c) The analysis has shown that offline pages in general, and the store locator in particular, are highly relevant for transactions, particularly for customers with a preference for pick-up in store. We found that nearly one-fourth of all Web site visitors in our sample accessed the offline concept, which indicates the importance of physical stores to a Web site. (d) Lastly, our results indicate that not all visitors accessed the site via the home concept. Thus, the Web site should further analyze how visitors access and browse the site in order to identify the most profitable referrers and navigation paths.

3.5 Session cluster analyses

This section proposes a set of Web analyses that groups online visitors according to their interests, as evidenced by their browsing behavior. The results are useful to determine and segment users’ browsing behavior in order to improve site design and to derive information about a site’s success in attracting specific groups of visitors.

We distinguish three types of clustering approaches depending on the data used:

Single-session clustering Different clustering techniques have been applied on user sessions: k-means [Mobasher, et al., 2002; Shahabi, et al., 1997], hierarchical clustering using concept hierarchies to describe visited pages [Fu, et al., 1999], or more encompassing descriptions to create user profiles [Heer and Chi, 2002; Mobasher, et al., 2000b].

Multi-session clustering By taking the set (or sequence) of all accesses associated with one cookie instead of the set (or sequence) of all accesses within one session, the basic unit of analysis again becomes the user. It can be expected that knowledge about multiple sessions of single users on the same site could lead to a number of valuable insights; every follow-up session of a single user could be used to confirm users’ interest in that information section. However, a repeat visit could also mean that information was not found. Furthermore, cookies reidentify visitors, not individuals. The predictive value of such information should therefore not be overestimated.

Transaction Clustering By adding demographic data about a user as further variables to the feature vector defined by that user’s navigation, further insights could be gained. Promising candidates for an analysis of multi-channel behavior include transaction

preferences (offline pick up, online payment, returns to stores, etc., cf. Section 3.3.2), or demographic data such as income. The combined analysis can provide useful insights into consumer preferences, as the example in the following section demonstrates.

3.5.1 Transaction clusters

We analyzed session clusters for the two transaction groups of online customers, one preferring direct delivery, the other pick-up in store (cf. Section 3.4.4). By again investigating their visits to the different concepts, we derive information about specific user profiles. Using k-means, we clustered the two groups of purchase sessions that have a preference for direct delivery and pick-up in store.

We obtained five clusters, each as shown in Table 3-7 (a) and (b).

(a)

Cluster	*1*	*2*	*3*	*4*	*5*
Home	2	1	2	2	2
Infocat	7	2	4	23	16
Offinfo	3	0	1	2	0
Infprod	6	3	12	21	5
Service	10	3	2	4	4
Transact	6	2	3	4	4
Number of cases	29	188	45	15	37

(b)

Cluster	*1*	*2*	*3*	*4*	*5*
Home	1	4	18	1	4
Infocat	22	30	6	1	6
Offinfo	1	7	1	5	19
Infprod	1	27	5	22	8
Service	5	4	0	0	12
Transact	3	7	3	3	4
Number of cases	25	15	147	40	55

Table 3-7: Cluster centers of weighted-concept purchase sessions with (a) direct delivery preference and (b) pick-up in store preference

Table 3-7 (a) shows visitors who chose direct delivery. They tend to be “true online users” (all clusters tend to rarely visit the offline concept). They fall into five subgroups: the largest group (cluster 2) tends to visit all other concepts except offline information. The number of page impressions is small. Groups 3, 4 and 5 tend to visit the semantically related concepts *infcat* and *infprod* and can be characterized as typical information seekers [Moe, 2001]. A small group (cluster 1) focuses on service-related information and exhibits the highest number of page impressions in this cluster group. The results are highly significant with $p < 0.0001$. Twelve sessions have been eliminated due to outlier sensitivity in k-means.

Table 3-7 (b) shows visitors who picked up their purchase in-store. They tend to be ”true multi-channel users” (nearly always visiting the offline concept). Its largest subgroup (cluster 3) takes advantage of all the site’s information offers and visits the offline concept at least once. A smaller subgroup (cluster 5) appears to be arriving with prior knowledge of their intended product choice; they do not need to consult the catalog or refer to service pages extensively but move directly to the service, offline and transaction concept. This may be interpreted as showing that these users combine the wish for a fast transaction process (online) with the reassurance that because they will pick up the product in-store, problems that may surface can be solved then. Clusters 1, 2 and 4 all focus on the concepts *infcat* and *infprod* before they move to the transaction concept. The results are highly significant with $p < 0.0001$, with the exception of the home concept ($p < 0.15$).

Similarities in the information behavior exist between cluster group 1 (pick-up) and group 2 (direct delivery). Cluster 1 in group 2 and cluster 5 in group 1 look at many catalog sites before moving to the transaction process; cluster 2 in group 2 and cluster 4 in group 1 both intensively explore information catalog and product information pages; cluster 4 in group 2 and cluster 3 in group 1 primarily look at product information.

3.5.2 Summary and implications

The presented clustering method demonstrated how user groups can be segmented based on Web usage data and how Web user data can further enrich the analysis. The analysis found several session clusters exhibiting a distinctive interest in offline information. These clusters indicate groups of site visitors that use traditional channels for purchases. The analyses are useful for Web marketing [Moe, 2001] and for Web applications such as recommendation engines or personalization systems that require a model of user behavior, which will be discussed in more detail in Chapter 5. Site managers can also use the analysis results to make the online presence more appealing to most profitable target groups. For example, site managers could improve the links between Web pages that are visited together. Our transaction clusters support the identification of those sets of pages that may lead to a purchase.

3.6 Demographic and order analyses

This section of our analysis framework will present a set of Web analyses for customer segmentation based on demographic and order characteristics.

Section 3.6.1 calculates the distance-to-store metric which measures the distance between customers’ zip code locations and the nearest store of the retailer and compares it with the purchase proclivity. The results can be useful to determine a Web site’s success in attracting new online customers, to determine places for new shop openings

and to investigate cross-channel effects between online and offline sales channels.

The second set of analyses focuses on the question of a customer's value to a company. Section 3.6.2 introduces the revenue concentration and the Gini coefficient, which analyze the cumulative revenue generated by a cumulative proportion of customers. Section 3.6.3 introduces an index of customer value, which is based on the purchase variables *frequency*, *recency* and *monetary value*.

The analyses are calculated based on transaction data from the multi-channel retailer and on demographic data that has been acquired from Deutsche Post Direkt [Deutsche Post Direkt GmbH, 2004].

3.6.1 Distance-to-store distribution

This section investigates whether the distance from an online customer's zip code location to the nearest physical shop has an influence on purchase proclivity. Two outcomes appear plausible: people who live farther away from a shop may have the same probability of becoming an online customer (easily substituting visits to physical stores for online purchases), or they may have a lower proclivity to purchase online (possibly because of a lack of trust in an online-only retailer). A third, though unexpected, option is that they may have a higher proclivity to purchase online. To obtain answers to these questions, a data set of online customers with home addresses that are distributed across the country is needed.

Our sample of 13,653 online customers was spread over an area of approximately 80,000 square kilometers (km^2). Data was acquired that links a zip code area to a longitude/latitude value. The zip codes included an area of x_{av} = 43 km^2 on average with values ranging from 2 to 200 km^2. For most countries, geographical data is also available on a more fine-grained basis such as on street and household level. However, for the purpose of a first approximation and demonstration of the measuring technique, five-digit zip code data was regarded as sufficient to match geographic coordinates with a customer's location.

We therefore investigated this question by analyzing the larger sample of 13,653 customer records. Distance to the nearest store was calculated as follows: it was assumed that (a) customer, shop, and population are located at the center of their respective zip code areas; (b) home address and shipping address were identical

(negligible error[13]); and (c) the online purchasing probability is equally distributed among the population.

We then calculated minimal distances between customer zip code and shop zip code[14]. The mean distance was x_{min} = 10.01 km with a standard deviation of s_{min} = 9.32 km. For the number of customers per zip code area, it was found that x_{cus} = 2.98 with s_{cus} = 2.81.

The mean population density for zip code areas was x_{pop} = 12,469 with a standard deviation of s_{pop} = 58,891. Then the correlation was measured between the number of customers from each zip code area – normalized with the respective population density in each zip code area – and their distance to the next shop.

Thus, let x be the number of online customers divided by the number of inhabitants in a given zip code area, n and y be the distance to the next store, then the distance-to-store correlation r_{dst} can be calculated as

$$r_{dst} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}.$$

Figure 3-9 shows that the larger the distance of a region to the nearest shop, the fewer customers this region contains.

[13] Shipping and billing address were identical for 94% of the customers with delivery preferences. More than two-thirds of the customers specified that their billing address is their home address. One-third refused to provide this information. Most of the customers who preferred to pick up orders chose the store closest to their contact address.

[14] MIN [D(km) = ARCCOS (SIN (Latitude CustomerZIP * PI / 180) * SIN (Latitude ShopZIP * PI / 180) + (COS (Latitude CustomerZIP * PI / 180) * COS (Latitude ShopZIP * PI / 180) * COS ((Latitude ShopZIP - (Longitude CustomerZIP)) * PI / 180))) * 6370 (=earth radius in km)]

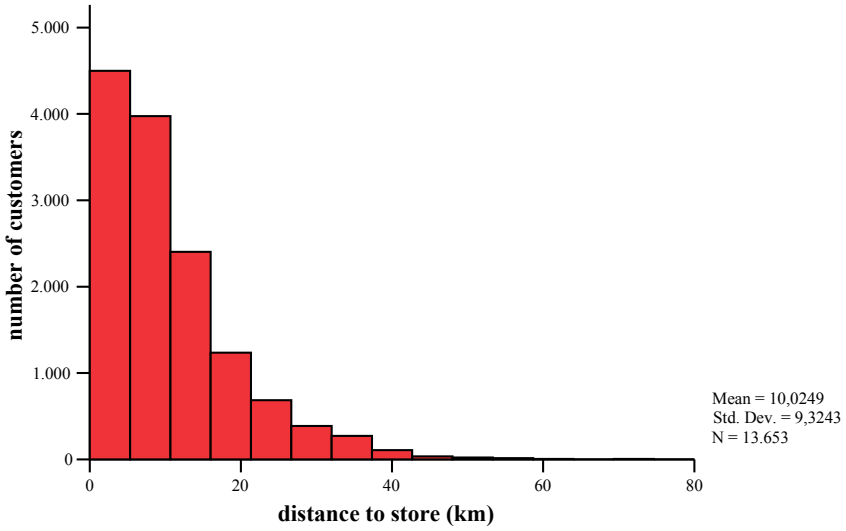

Figure 3-9: Histogram displaying the number of online customers and distance to store

We found a weak correlation of $r = -0.3$ and $p < 0.001$. This result could be an artifact if regions that are farther away from a shop (e.g., rural regions) simply contain fewer residents. However, in comparison, this relationship between population density in a zip code area and the next shop is so weak ($r = 0.01$; $p < 0.001$) as to be practically meaningless. That is, the presence of a physical store in one's vicinity appears to heighten the probability of shopping online with that company. What effects does the vicinity of a store have, then, on transaction preferences? There is indeed evidence of the expected relationship: customers from the all-customers sample who picked up their purchases in-store ($n = 9073$) lived, on average, 7.87 km from the nearest branch, while those who chose direct delivery ($n = 4580$) lived, on average, 12.15 km away. This relation was also mirrored in our online sample (average distance of direct-delivery customers from the nearest shop, $n = 621$: 13.01 km). Delivery preference, in turn, can be linked to Web usage behavior, as we have seen above. The geographic distribution of stores and customers has been depicted in Figure 0-3 of the Appendix.

The results are consistent with [Kohavi, 2003], who found that people who live farther away from retail stores spend more on the average and account for most of the online revenues. Our results are also consistent with the findings of the multivariate analysis of user perceptions in Chapter 2 where online consumers' trust in an e-shop has been influenced by perceived size and reputation of a retailer's physical presence.

Summing up, a Web site must cater to the needs of those prospects who need to rely on direct delivery, in particular by providing adequate information about the company, the products and transaction terms in its service pages. Besides this rather evident conclusion, a site could use the geographical findings as an indicator for the site's

success in attracting new customers through the Web. Consumers who live far away from the next shop are less exposed to physical stores and more likely to purchase online. Finally, the findings could be used to determine places for new shop openings in order to utilize the observed cross-channel effects between the Internet and a small-meshed store network. Combined with information about the offline conversion rate it may encourage companies to further integrate their online and offline marketing.

3.6.2 Concentration indices

A Web retailer must generate revenue to be successful. Thus, one of the most important segmentation criteria is the revenue contribution of customers. A Web site should cater considerably to the needs of those customers who generate the highest revenue.

In order to find out if there is a group of customers with a high revenue contribution, the Lorenz curve can be drawn, which is a useful method to depict, calculate and compare the revenue concentration in a customer sample. The Lorenz curve is defined as the function of the cumulative proportion of ordered individuals in subsets mapped onto the corresponding cumulative proportion of their size [Lorenz, 1905].

Given a sample of *i* ordered customers with the revenue *r* respectively, then the Lorenz curve can be expressed as

$L(i) = \sum_{k=1}^{i} r_k$. In the case of the multi-channel retailer, the Lorenz curve revealed that 20% of the retailer's customers generate 60% of the revenues. Though the often cited Pareto rule that 20% of customers typically generate 80% of revenue [Koch, 1998] could not be fully confirmed, a tendency towards revenue concentration could be observed.

The Gini coefficient is a summary statistic of the Lorenz curve and a measure of inequality in a population. The Gini coefficient *G* is defined as

$G = 1 - \sum_{i=1}^{n} (\partial Y_{i-1} + \partial Y_i) \times (\partial X_{i-1} - \partial X_i); 0 \le G \le 1$, where ∂Y_i and ∂X_i are cumulative percentages of X_i, the population variable, Y_i the income (or revenue) variable and *n* the number of observations. *G* ranges from a minimum value of zero (total equality) to a theoretical maximum of one (total inequality). In the sample of 13,653 online customers at the multi-channel retailer, the Gini coefficient was $G = 0.41$.

3.6.3 Recency, frequency, monetary value

The question arises if revenue is a reliable indicator to determine a customer's value to the company. Is a one-time customer who spends a lot in a single transaction more

valuable than a customer who spends less but more frequently on a long-term basis? Further purchase characteristics need to be examined to segment customers according to their value to a company. A typical index for determining customer value is based on three variables: the time of the most recent purchase (*recency*), the number of orders placed (*frequency*) and the amount of money spent[15] (*monetary value*) within a specific time frame [Miglautsch, 2000].[16]

In order to calculate the index, the following scores have been assigned to the three purchase characteristics:

Score	*Recency of last purchase*	Score	*Frequency of purchases*	Score	*Monetary value*
1	> 6 months ago	1	one per year	1	< 200 euros
2	3 to 6 months ago	2	2-3 per year	2	200-600 euros
3	< 3 months	3	> 3 per year	3	> 600 euros

Table 3-8: Recency, frequency and monetary value scores

Customers were grouped according to their purchase characteristics. In total, 27 segments (3x3x3) were generated from the score combinations. For example, the segment with the score code 312 contains all customers whose last purchase took place more than six months ago, who purchased more than three times, and whose total purchase value was between 200 and 600 euros.

Customers with the same points in all categories were grouped and the results depicted in Figure 3-10. The abscissa is partitioned into 27 segments which are assigned the number of customers that belong to this class.

[15] Often profitability is used instead of revenue.

[16] Recency and frequency have been used also in the context of Web site visitors [Cutler and Sterne, 2000]: Visit recency measures the time of the most recent visit and visit frequency the number of visits in a time frame.

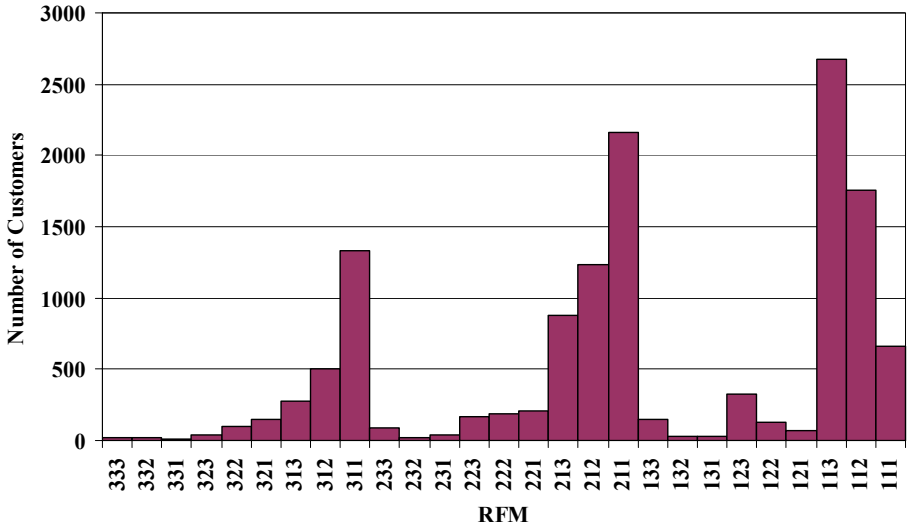

Figure 3-10: Recency, frequency, monetary value distribution for 13,653 customers

Segments 113, 211, 112 and 311 contain the most records. These segments rank lowest (1) in at least two variables. Only very few customers rank highest (333) in all three variables. The retailer should subsequently focus its business efforts on the needs of those segments with the highest scores in all three variables. One should note that the data sample of 13,653 customers in this analysis includes purchases from a time period of just eight months. The results will be different for longer time periods. Within the given time frame, the mean transaction amount per order was 672 euros, the mean number of purchases per customer, 1.14, and the mean interpurchase time between two consecutive orders of the same customer, 156 days.

The presented analysis is popular for customer segmentation due to its simplicity. Criticism concerns the creation of equal bins [Miglautsch, 2000]. More fundamental criticism aims at the variables used to determine customer value. Reinartz and Kumar [2003] compared transactions from more than 11,992 households at a catalog retailer over a three-year period and found that scoring approaches resulted in an overinvestment in advertising cost for lapsed customers.

3.6.4 Summary and implications

We demonstrated how users can be further segmented according to demographic and order characteristics.

The distance-to-store analysis, which indicates the site's success in attracting new customers through the Web has been calculated. The findings could be used to determine places for new shop openings in order to utilize the observed cross-channel effects between the Internet and a small-meshed store network. Moreover, the correlation

provides insight into the potential relevance (and potential explanatory value) for different service choices in multi-channel retailing.

The concentration indices provide a better understanding of the customers' revenue contribution to a company's business success. A customer value index has been suggested that measures the value contribution of distinct customer segments.

The results can be also useful for recommendation and personalization systems [Kobsa, et al., 2001; Sarwar, et al., 2000].

3.7 User typology analyses

This last section of analyses within our framework will introduce a method of pattern discovery that allows the identification of user typologies expressed as browsing strategies. This notion of success is particularly useful for information Web sites where a site's goal is to attract specific types of online visitors and to keep them recurring to the site.

Section 3.7.1 discusses the notion of success for an information site. Section 3.7.2 introduces how behavioral strategies can be modeled on Web usage data. Section 3.7.3 discusses how these strategies can be expressed in a Web mining language. Section 3.7.4 describes the information Web site, and Section 3.7.5 introduces a concept hierarchy for that site. Section 3.7.6 demonstrates how a specific behavioral strategy could be tested against Web usage logs from the information Web site. Section 3.7.7 presents the results and discusses the discovered patterns.

3.7.1 Success for an information site

The presented analyses from the previous sections consider user behavior in the context of Web merchandizing. However, the Internet contains an abundance of non-merchandizing sites, in which a similar behavior should be expected. On an information site, objectives of the interaction may be the retrieval of pages on a subject of interest: the enrollment in a course, the identification of an appropriate partner or the application for a job. Thus, success may have different meanings depending on the site's goals. Events such as filling out a registration or application form, downloading information, ordering a newsletter, the use of a product configuration tool, signing a contract or contacting a physical person may define conversion success in a non-merchandizing context. This chapter will introduce a method how success can be determined on an information Web site.

In the following, we apply a Web analysis methodology on the Web log data of a non-merchandizing site. The data owner belongs to the category of organizations that use the

Web mainly as a contact point, in which visitors are motivated to a face-to-face contact. Thus, this category encompasses sites of sophisticated services, including Application Service Providers (ASPs), insurance companies and consultancies, as well as companies offering personalized customer support. In the absence of cookie identifiers, sessions were determined heuristically [Berendt, et al., 2001; Berendt and Spiliopoulou, 2000; Cooley, et al., 1999] specifying 30 minutes as a threshold for viewing a single page of a session. After cleaning and preprocessing, the cleaned server log contained 27,647 user sessions.

3.7.2 Modeling strategies as sequences of tasks

The process of becoming a customer has been described for e-commerce sites in Section 3.3.1 where the purchase process has been used as a model for site design and for the interpretation of the behavior of potential customers. This task-oriented view on browsing behavior can be useful in the context of information Web sites, too.

More generally, we define a "strategy" as a sequence of tasks, beginning at a start-task, ending at a target-task that corresponds to the fulfillment of the objective of the interaction, and containing an arbitrary number of intermediate tasks.

Hence, if we observe the set of conceivable tasks in an application as a set of symbols *S*, a strategy is a regular expression involving at least two symbols from *S* (the start-task and the target-task) and, optionally, a number of wildcards. Borrowing from the conventions on regular expressions upon strings, we propose the following notation for the representation of strategies:

- *A strategy is a sequence of symbols from the set of tasks S, optionally interleaved with an arbitrary number of associated wildcards.*
- *A wildcard has the form [n;m], where n is a non-negative integer, m is a non-negative integer or a symbol denoting infinity, and n ≤ m.*
- *A wildcard [n;m] appears as suffix to a task or a parenthesized subsequence of symbols, indicating that this task or subsequence should occur at least n and at most m times.*

The first and the last element of a strategy and of any subsequence suffixed by a wildcard are tasks from *S*, i.e. they may not be wildcards.

The first task or subsequence of a strategy may be prefixed by a special symbol # indicating that this task is the very first occurring in data records conforming to the strategy.

Similarly to string matching for regular expressions, a strategy is matched against sequences of events from the dataset. In Web usage mining, these sequences are user sessions derived from the Web server log [Cooley, et al., 1999].

3.7.3 Expressing strategies in a mining language

The specification of a strategy according to the notation used in the previous section is appropriate for strategy generation. However, in order to discover patterns adhering to an anticipated strategy, we must express a strategy formalized in a mining language.

Our method of pattern discovery uses the specification of the behavioral strategy itself as guidance to the analysis software. Findings from cluster analysis or association rule mining (cf. Section 3.3.3 or Section 3.5.1) can be used as guidance for the strategy specification.

Hence, the challenge lays in modeling the behavioral strategies of users in such a way that they can be tested against Web usage data.

To this purpose, we use the Web mining language MINT of WUM (Web Utilization Miner) [Spiliopoulou, 1999; Spiliopoulou and Faulstich, 1999].

In MINT, a strategy is mapped onto a template. A template is similar to a regular expression, comprised of variables and wildcards. A task that should appear in a strategy corresponds to a bound variable. A wildcard in a strategy is directly mapped into a wildcard of the template. The constraints for the first and last elements of a strategy are also valid for templates.

During data mining, templates are matched against groups of sessions: a session matches a template if it contains all tasks contained in the template in the appropriate order and, further, satisfies all constraints posed by the template. In the context of strategy evaluation, strategies express the *expected* behavior of users, while sessions reflect the actual behavior recorded in the Web server log. Thus, a session is “conformant” to a strategy if and only if it matches the template expressing the strategy.

3.7.4 An informational Web site

The Web site of the case provides information material and contact points on several services. Visitors access the site to be informed about the company, its mission and profile, its product portfolio, its credentials, partners and reference customers. Conversion corresponds to the initiative of the visitor to contact or become contacted by the company. In some sites, the execution of a “Contact” task is a unique event during a session: the user provides her contact data, so that a meeting can be arranged. In other sites,

including the one at hand, a contact task may be the acquisition of information material on a given product or the registration to an event organized by the company. In such a case, a contact task may be executed multiple times, once per product or event of interest. Hence, a session may contain multiple "Contact" task invocations.

Its users include potential members, actual members, institutional partners, personnel and press. For the purposes of the analysis, we have concentrated on the behavior of potential members and have removed all sessions that could be identified as belonging to actual members or personnel, as well as visits of robots, archivers and administration services, which are identified by their IP address. Invocations of components of each individual page (tables, images, script invocations) were coerced into a single page view by a site expert.

3.7.5 Task-based site taxonomy

Figure 3-11 shows the task-based taxonomy of the Web site. The *service* pages provide primarily information for existing customers, including services and responsible contact persons. The *research* pages contain information about important projects and relevant reports. Of special interest for our study is the branch under *marketing/public relations (PR)*. Here we aggregated all pages primarily dedicated to marketing purposes. Information pages under *acquisition* contain detailed information of programmes offered by the organization. Pages providing online registration forms, detailed contact data or downloads of application material were summarized under *registration*.

For the given information site, the conventional process of the customer purchase process must be replaced by a reasonable sequence of tasks modeled in the concept hierarchy. In our example, "Conversion" corresponds to the establishment of a contact, i.e. to the execution of a "Contact" task according to Figure 3-11.

Figure 3-11 also shows how each concept was assigned to one out of the three phases of the online information process consisting of background information, detail information and contact. The registration pages were assigned to the *contact* phase, while the acquisition-related information pages were mapped onto the *detail information* phase. All remaining pages where treated as providing *background information*.

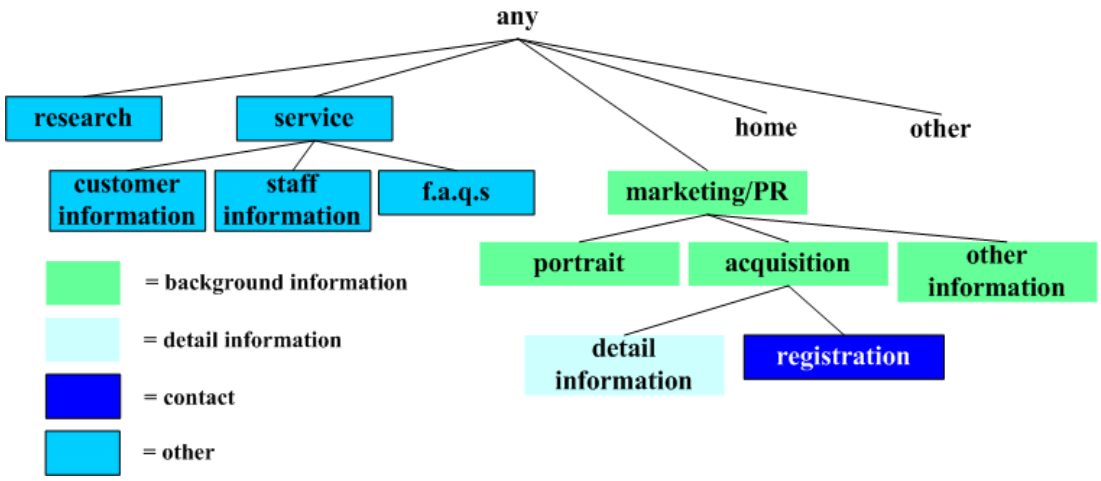

Figure 3-11: Task-oriented taxonomy of the information site

3.7.6 Mining queries for template matching

This section shows an example how a behavioral strategy – namely the knowledge building strategy proposed by Moe [Moe, 2001] – could be tested against Web usage logs. The study of background information, corresponding to the invocation of the "BackgroundInfo" task in the taxonomy above, is expected to characterize the *knowledge builders* [Moe, 2001]. These users prefer to get the complete picture of the company, to check the mission and verify the trustworthiness of the institution, before deciding to establish a contact. Background information may be acquired before or after executing a "DetailInfo" task. As the behavior of these users cannot be traced beyond a single session, we have rather concentrated on a subgroup of knowledge builders, namely those that acquire enough information about the company *and* establish a contact within the same session. It should be noted that Moe's model cannot be applied in its entirety, because it contains strategies that are only relevant for e-commerce sites.

According to the task-oriented taxonomy of Figure 3-11, the knowledge building strategy has the form:

 # Home (BackgroundInfo[1;n] DetailInfo[1;n])[1;n]

We use the mining language MINT to extract the pattern for the templates of the knowledge-building strategy. MINT is an SQL-like mining language for the specification of templates and of constraints upon them. The full syntax of MINT is presented in [Spiliopoulou and Faulstich, 1998].

```
SELECT t

FROM NODE AS x y z w, TEMPLATE # x y * w * z AS t

WHERE x.url = "Home" AND y.url = "BackgroundInfo"

AND wildcard.w.url = "BackgroundInfo"

AND w.url = "DetailInfo"

AND wildcard.z.url ENDSWITH "Info" AND z.url = "Contact"
```

Table 3-9: Strategy specification in MINT

The template expresses the strategy as a sequence of variables and wildcards. The first three constraints bind the variables. The last constraint binds the contents of the wildcard.

3.7.7 Results and analysis of the discovered patterns

The navigation pattern returned a group of paths. Each task in each path has been invoked by a number of visitors, some of which followed the path to the end, while others have abandoned it. In our case, these are the routes from “BackgroundInfo” to “DetailInfo” and then to the invocation of the “Contact” task. All these paths consist of “BackgroundInfo” and “DetailInfo” tasks. However, one visitor may have invoked “DetailInfo” after asking for “BackgroundInfo” once, while another may have requested “BackgroundInfo” ten times beforehand. Moreover, each path has been entered by a number of visitors, some of which have followed it to the end, while others have abandoned it.

The invocation of detailed information indicates a serious interest in the offered product or service. Hence, we split the pattern of this strategy into two components, one until the first invocation of “DetailInfo” and one thereafter. The statistics of the first component of the knowledge-building strategy are shown in Figure 3-12. The horizontal axis represents steps, i.e. task invocations. At each step, a number of users asks for detailed information and thus proceeds to the second component of the strategy. These users are represented in the cumulative curve labeled “DetailInfo”. The vertical axis shows that from the 6,641 visitors that entered this strategy, about 14% (896 visitors) entered the second component. The remaining ones are depicted in the cumulative curve labeled “Exit”: they did not necessarily abandon the site, but their subsequent behavior does not correspond to the knowledge-building strategy any more. The curve labeled “BackgroundInfo”, represents the visitors that ask for further background information. All curves saturate fast, i.e. most users invoke only a few tasks.

The statistics of the second component are shown in Figure 3-13. After the first invocation

of “DetailInfo”, the 800 visitors that entered the second component acquired detailed or background information aggregated into the curve “Info” that covers invocations of both tasks. The “Contact”-curve and the “Exit”-curve are again cumulative. The former shows that 10% of these visitors establish contact, and that they do so after a small number of information acquisition tasks. This implies that many contact acquisition tasks do not increase the confidence of contact establishment. The large number of users represented by the “Exit”-curve indicates that the strategy does not represent all users. Hence, further tasks should be modeled and more strategies should be investigated.

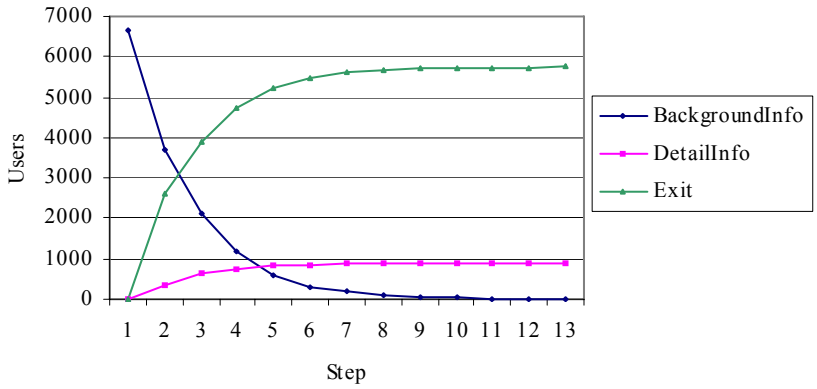

Figure 3-12: Knowledge-Building Strategy until the first invocation of “Detail Info”

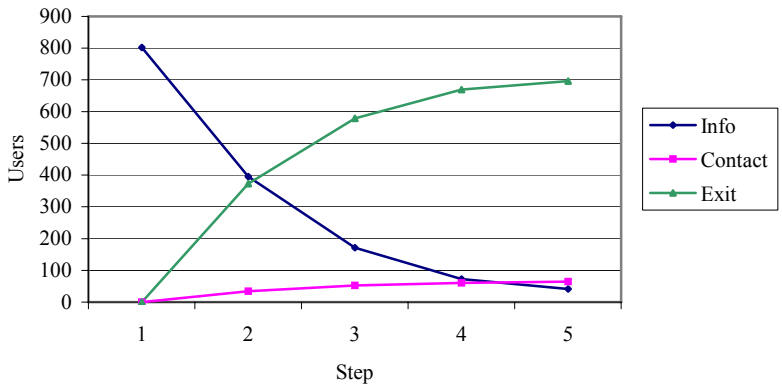

Figure 3-13: Knowledge-Building Strategy until contact establishment

3.7.8 Summary and implications

We addressed the issue of analyzing Web site usage according to the anticipated goals of site visitors. To this purpose, we have presented an approach for the modeling of user activities as tasks in pursue of a goal, and of sequences of tasks as strategies to achieve this goal. Our framework allows for the description of navigation strategies as anticipated

in marketing literature. However, our model is not limited to Web merchandizing. We have demonstrated the applicability of our approach by analyzing the behavior of two types of visitors on an information Web site.

The modeling of goal-oriented navigation strategies is a non-automatable task. However, the specification of appropriate constructs for the formulation of strategies is essential. The current framework and the mining language we use for the discovery of patterns adhering to a strategy are a first step in this direction.

3.8 Conclusion

Five groups of Web analyses have been presented that constitute our analysis framework. The group of service analyses in Section 3.3 is beneficial for multi-channel retailers in order to determine consumers' delivery, payment and return preferences. The conversion metrics in Section 3.4 analyze consumers' navigation behavior on a fine-grained level. The offline conversion rate can be used as an indicator for the site's success in inducing offline sales. The clustering method presented in Section 3.5 is useful to improve Web site navigation and to identify navigation patterns of online buyers. Section 3.6 presented order and demographic analyses that group users according to demographic and order characteristics. The distance-to-store metric has been defined that indicates the site's success in attracting new customers. The proposed customer value indices provide a first insight in a customer's value contribution. The analysis of user typologies in Section 3.7 modeled user activities as sequences of tasks. The method allows searching for specific user navigation patterns in the Web log.

The results of our Web analysis framework should be compared over time. A comparison is beneficial for tracking how modifications of Web site design, product and service offerings or advertising may influence the analysis results and Web site success respectively. Moreover, a company can use the results to identify trends and patterns over time in order to predict future demand in Web site content, services and products.

The clustering results of Section 3.5, the order and demographic characteristics of Section 3.6 and the user typologies of Section 3.7 are particularly useful for user modeling in personalization systems, which will be discussed in more detail in Chapter 5. The importance of Web mining for personalization has been described in related work [Mobasher, et al., 2000a; Mulvenna, et al., 2000; Perkowitz and Etzioni, 2000; Spiliopoulou, 2000]. Personalization systems need to acquire a certain amount of information about users' interests, behavior, demographics and actions before they work efficiently. As multi-channel retailers can collect consumer information from several distribution channels, personalization can be particularly beneficial for these retailers.

3.9 Limitations

This chapter has used data samples from a retail and an information site to test the metrics and analytics of our analysis framework. However, site-specific parameters could limit the generalizability of the results. It could be that the specific structure of the Web sites or the products and services offered have an impact on the analysis results. Thus, if a company wants to compare its performance with other sites, site-specific criteria need to be included in the discussion of the results. As our sample of customers and Web logs is relatively large it could be used for comparisons with other sites.

For the development of conversion metrics and the modeling of user search strategies we referred to the purchase decision process, which is a well-known model of consumer purchasing behavior. However, decision processes could be more complex in reality, which may not be captured by the proposed analyses.

The list of 82 metrics and analytics (cf. Table 0-3) is a selection of analyses that covers important aspects of success measurement and customer relationship management for Web sites. It was considered useful by experts and the Web site owners. Of course, the selection is not exhaustive and can be further expanded.

4 Prototypical development of a privacy-preserving Web analysis service

Companies' data collection and analysis practices as described in Chapter 3 have increased users' privacy concerns significantly, which is a major impediment for successful e-commerce. Privacy legislation has been implemented in many countries to alleviate some of these concerns. Moreover, site owners are increasingly adopting an industry standard for privacy protection – the Platform for Privacy Preferences (P3P) – that gives users more control over their personal information when visiting Web sites. The implications of these privacy requirements for our analysis framework from Chapter 3 will be discussed. This chapter will present a prototypical Web analysis service that calculates the analyses of Table 0-3 and indicates respective privacy requirements.

The chapter structure follows the main phases of the software development process [Sommerville, 2004]. Section 4.1 presents the main idea of the prototype's business model. Section 4.2 concentrates on privacy requirements and their implications for the calculation of metrics and analytics in our analysis framework. Section 4.3 presents the prototype design which, given a set of privacy constraints and available data elements, selects the Web analyses that are allowed to be calculated. The main functions and processes are presented. The specification of constraints arising from the specified privacy requirements is formulated as a syntactical extension to P3P. Section 4.4 presents the user interface. The main selection parameters and output formats are described. Section 4.5 discusses the implementation of the prototype.

Section 4.6 will briefly discuss how disallowed analyses could be modified in such a way that they return altered but still useful results without comprising privacy requirements. The goal is to reach a maximum amount of privacy to the customers while still allowing the site analyst to obtain valuable results.

4.1 Business model

The main function of our privacy-preserving analysis prototype is to calculate those Web analyses in our framework (cf. Table 0-3) that are not restricted, given a set of privacy constraints and data elements. Moreover, if a site is P3P-enabled the analysis service automatically parses the specifications about available data, purpose and jurisdiction and indicates potential restrictions when metrics and analytics are calculated.

The Web site owner can be located in any country. However, legal privacy restrictions are currently only specified for German retailers. If the site is not P3P-enabled manual specifications are required. The Web service business model is depicted in Figure 4-1:

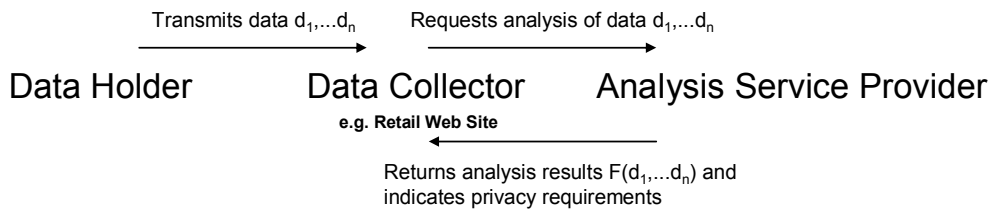

Figure 4-1: The Web Service business model

Data can be exchanged by electronic transmission (e.g. by download) or on a physical storage medium. If the data collector has trust in the service provider and legislation does not restrict the use of certain data for analysis purposes, the complete data set can be transmitted without modifications. However, in the unlikely case that the analysis service is mistrusted (if so, it would be unlikely that the two parties engage in a business relationship) or legal policies restrict the use of certain data for analysis purposes, the retailer must protect confidential information before the data is transmitted. Methods to protect sensitive data are discussed in Section 4.6.

Note that the tool does not protect the consumer from deliberate privacy violations by the retailer or the service provider. It only supports the data analyst in calculating allowed analyses and recognizing potential privacy conflicts and possible usage purposes that must be respected. Thus, the business model requires that all parties must be trusted.

Standards for secure communication, e.g. a Secure Socket Layer (SSL) [Stallings, 1999], are integrated in the framework. Further security questions such as attacks from a malevolent hacker or employee are not the scope of this work.

The Web service could be offered for a per-service fee or as a renewable or permanent license. The business model could be enhanced by comparing analysis results between companies to create and sell benchmark reports for specific industries. In this case, further privacy measures have to be taken to protect shared data from misuse by third parties [cf. Boyens, 2004].

4.2 Privacy requirements

The following section will discuss privacy requirements and their implications for our analysis framework. Section 4.2.1 discusses privacy restrictions in German legislation. Section 4.2.2 presents the main specifications of P3P. Privacy problems from data inferences are discussed in Section 4.2.3. We will give examples of how inferences could bypass P3P specifications.

Implications from these requirements for the specification of our privacy-compliant analysis service are discussed and summarized in a problem statement in Section 4.2.4.

4.2.1 Legal restrictions

Laws protecting the privacy of individuals exist in more than 30 countries [Kobsa, 2002]. A number of regional, industry-specific and transnational regulations have been adopted in addition. It is beyond the scope of this thesis to discuss and compare privacy laws in different countries and their national and transnational implications in detail. Comprehensive resources are available for this purpose [e.g. www.epic.org, www.privacy.org, www.privacyinternational.org, www.privacyexchange.org, Agre and Rotenberg, 1997; Andrews, 2002; Rotenberg, 2001]. The legislative requirements in this section will focus on German privacy laws.

Directive 2002/58/EC of the European Parliament and the Council concerning the processing of personal data and the protection of privacy in the electronic communications sector [EU, 1995] and its extension for electronic data [EU, 2002] has been adopted in national laws in most European Union (EU) member states [EU, 1995]. In Germany, the EU Privacy Directive has been implemented in the Federal Data Protection Act [BDSG, 2003] and in the privacy laws of the German states [cf. EU, 1995]. For electronic services such as e-shops the Teleservices Data Protection Act [TDDSG, 2001] contains further, more specific regulations. TDDSG and BDSG regulate the collection, processing and usage of person-related data (§1 (2) BDSG, §1 (1) TDDSG). Person-related data is defined as information about identified or identifiable[17] persons (§3 (1) BDSG). The TDDSG differentiates between "stock" data that is necessary for the reasons, contextual form and changes of a contractual relationship (§5 TDDSG) and "usage" data that is required for the usage of services (§6 (1) TDDSG). §3a BDSG imposes an obligation to collect data only in a sparing and avoidable way. A more detailed discussion of legal implications for e-commerce in Germany can be found in Hansen [2002].

The following sections discuss the main implications of German privacy laws on the analysis of user and usage data in our analysis framework. The main consequence of German privacy legislation for our analysis framework is that certain analyses are only allowed on pseudonymous data. Thus, analyses requiring identified data must be blocked by the analysis prototype. The requirements for the metrics and analytics are depicted in Table 0-3, where it is indicated whether an analysis requires identified or pseudonymous data. Privacy implications for three data types are discussed in the following sections for Web user data (4.2.1.1), Web usage data (4.2.1.2) and microgeographic data (4.2.1.3).

[17] When users can be identified with reasonable effort based on the data collected, privacy laws already apply.

4.2.1.1 Web user data

Data collected for billing purposes in electronic retailing is person-related (§5 TDDSG) and must be modified by the e-shop before it can be transferred to the analysis service.

In our cooperation partner's data schema (cf. Table 3-2), the (combinations of the) attributes `name`, `surname`, `street`, `street_number`, `recipient_address`, `e-mail_address`, `date_of_birth` and `phone_number` refer to identified or identifiable persons. Thus, all analyses that require attributes referring to identified or identifiable persons are disallowed according to German privacy legislation[18].

As indicated in Table 0-3 some metrics and analytics in the analysis framework require at least a pseudonymous recognition of customers. Legislation explicitly allows the creation of pseudonymous user profiles for analysis purposes (§6 (3) TDDSG). Thus, if an analysis requires pseudonymous user data, all identifiable attributes such as `name` and `surname` should be replaced by a pseudonymous `customer_id`. It should be noted that linkage of pseudonymous user profiles with other attributes may lead to reidentification of customers. This problem will be discussed in more detail in Section 4.2.3. The pseudonymization should be performed by a trusted party in the company (e.g. the data protection officer).

German legislation requires the deletion of identifiable transaction data not later than six months after the time of data collection (§6 (4), §6 (7) TDDSG). In order to perform pseudonymous analyses over a longer period of time, the company should establish two separate databases: a "business intelligence" database where only pseudonymous information is stored for data analysis and a "transaction" database where billing data is stored for order fulfillment. The analysis service should have access only to the business intelligence database.

A technical approach to incorporating privacy policy enforcement into an existing application and database environment has been proposed in Le Fevre et al. [2004]. Agrawal et al. [2004] proposed an auditing framework that determines whether a database system is adhering to its data disclosure policies.

---

[18] However, the visitor can give explicit consent to the use of her data for analysis purposes according to §3 (2), §4 (2) TDDSG. Moreover, individual analyses may be allowed if they are required for the fulfillment of the transaction purpose. Thus, it may be possible to compile a "black" list of customers who frequently did not pay.

4.2.1.2 Web usage data

Web logs are considered person-related because user sessions contain the attribute `ip_address` or other attributes possibly indicating a user's identity (e.g. `login_name`, `user_authentication`). In particular, users with a static IP address could possibly be identified.[19] According to §6 (3) TDDSG, a pseudonymous analysis of Web usage data would be possible. Thus, the data collector should perform the following pseudonymization steps before the data is transferred to the analysis service:

If an `ip_address` is required for session reconstruction, it must be replaced by a pseudonymous ID. Moreover, the e-shop must delete or pseudonymize all attributes possibly indicating a user's identity such as `user_login` or `authentication` before the log file is stored for analysis.

The `ip_address` is also required for the matching of localization and geographic information (cf. Section 3.1.1). This analysis would be illegal in German privacy legislation. The data collector could delete the last digits of the `ip_address`. However, this decreases the accuracy of IP localization tools significantly.

One should notice that the tables **`session`** and **`order`** could be combined via the attribute `access_time`, when the `customer_id` is assigned in consecutive time sequence. However, if all identity and identifiable attributes from the **`order`** table have been replaced by pseudonymous information, session data may remain anonymous and thus the analysis would comply with privacy legislation.

In order to recognize visitors in several sessions, cookies or login information are required [Berendt, et al., 2001]. The use of cookies, their settings and usage purposes should be explicitly communicated to the site users. The information stored in the cookie should not contain links to identified or identifiable information.

4.2.1.3 Microgeographic data

In contrast to the offline domain, where legislation has adopted a marketing-friendly privacy jurisdiction (cf. §28 BDSG), the TDDSG is more restrictive on the analysis and use of microgeographic data in the online domain. The combination of online billing

[19] However, telecommunication service providers can store a user's IP address and combine it with user data if it is required for billing purposes (§6 (2) TDDSG).

information and microgeographic data is illegal if a customer becomes identifiable [Weichert, 2004]. Thus, analyses that include online data in combination with microgeographic data are only legal if the user remains anonymous.

4.2.2 P3P specifications

Besides legal restrictions that are mandatory, companies can self-impose further restrictions on their data collection and data usage practices. A company's privacy practices are typically posted as online privacy statements (also known as "privacy policies" or "privacy disclosures").

A technical approach to codifying a company's Web privacy practices is the Platform for Privacy Preferences (P3P). It enables Web sites to express their privacy practices in a standard XML (Extensible Markup Language) format that can be retrieved automatically and interpreted easily by user agents. P3P is an industry-supported, self-regulating approach to privacy protection. It has been recommended by the W3C [Cranor, et al., 2002] as a protocol to communicate how a site intends to collect, use, and share personal information about its visitors. P3P adoption is 33% for the top 100 Web sites and 22% for the top 500 Web sites [Ernst&Young, May 2004].

P3P-enabled browsers parse a site's privacy policy automatically and compare it to the privacy preferences of the visitor, who can then decide to use the service or not. Once a P3P policy is set up on a Web site, it becomes a legally binding agreement predicated on notice and consent between the Web site and the user [Cranor, et al., 2002]. In the US, the Federal Trade Commission and several states have increasingly sued companies that did not adhere to their privacy policies for unfair and deceptive business practices.

P3P cannot constrain or modify existing privacy legislation. Thus, the use of P3P by itself does not constitute compliance with the EU Data Protection Directive, though it can be an important part of an overall compliance strategy [Cranor, et al., 2002]. The latest version of the P3P specification (Version 1.1 as of January 2005) includes a "jurisdiction" extension element where a known URL of a body of legislation can be inserted, which can be recognized by user agents.

The P3P 1.0 specification defines a base set of data elements a Web site may wish to collect, a standard set of uses, recipients and other privacy disclosures. A STATEMENT describes data practices that are applied to particular types of data. A STATEMENT element is a container that groups together a PURPOSE element, a RECIPIENT element, a RETENTION element, a DATA element, and optionally other information.

P3P is characterized by an "atomic" focus through its separate description of different

combinations of DATA, PURPOSE, RECIPIENT. This may lead to problems when data are combined. The implications of this problem for our analysis service are discussed in Section 4.2.4.

4.2.2.1 The DATA element of P3P

P3P provides a data schema built from a number of predefined data elements, which are specific data entities a service might typically collect (e.g. *last name* or *telephone number*).

The data schema in our privacy-preserving analysis tool parses the data elements specified in a P3P policy. Further data elements can be specified manually. Analyses are disabled if required data are not available.

4.2.2.2 The PURPOSE element of P3P

The description of the PURPOSE element requires site owners with P3P policies to explain and disclose the purpose of data collection for each DATA element or group that is collected. P3P suggests twelve standard purposes of data collection [P3P, 2002]. The PURPOSE specification for DATA elements does not restrict the calculation of analyses in our service tool. As discussed before, the use of the analysis results depends on the company's business interests and cannot be controlled by our analysis service. However, the service automatically indicates for what purpose(s) DATA elements were collected, which reminds a Web site owner to use the analysis results only for the specified purpose(s).

4.2.2.3 The RECIPIENT element of P3P

P3P STATEMENTS must include a RECIPIENT element containing one or more recipients of the data. In order to assure a legal use of the analysis framework, the Web site owner should specify that the collected data is received by the data collector (<OURS>) and the analysis service provider that uses the data under equable practices (<SAME>).

4.2.2.4 The RETENTION element of P3P

A STATEMENT element must also include information of the data collector's retention policy. In order to use the analysis service, the data collector should specify that data is retained for analysis purposes.

4.2.3 Inference problems

A problem that has not yet been directly addressed in P3P specifications is inferences from data that are re-combined after collection. Inferences[20] exploit the possibility of intersecting separate releases of identified and unidentified data. Even if identity keys are not known, attributes from secondary data sources may doubtlessly point out a single person [Denning, 1982; Sweeney, 2001].

Related problems have been described by the methods of *data and record linkage* [e.g. Fellegi, 1972; Newcombe, et al., 1992; Winkler, 1995], *pattern matching* with aggregation operations [e.g. Torra, 2000] and *object identification* [Neiling, 2004].

Sweeney [2002] used publicly available information from a voter's registry containing the data attributes `name, age, gender` and `address`. These attributes were compared with "anonymized" patient records from hospitals (where patient names had been deleted). Sweeney found that the attributes {`date_of_birth, 5_digit_ZIP_code`} identified 69% of the patients, {`date_of_birth, gender`} identified 29% and {`date_of_birth`} identified 12%.

Inference problems are also given for geomarketing, where customer attributes such as `customer_id, gender, data_of_birth, credit_rating, zip_code, street_name, street_number, pages_visited, product_name` are exchanged and matched with secondary demographic data. The matching would be legal in the online domain if the user profile remains pseudonymous. Deleting name and address would be a first step towards pseudonymization. An analysis would become privacy-critical, however, if a `customer_id` and `zip_code` are linked to the `product_name` ordered. In this case conclusions could be drawn from the customer's preferences and residence. For example, a researcher who orders specialized books in his field of interest is likely to be identified by the zip code that indicates the location of his university or research institution. Especially for sparsely populated zip code areas – the smallest zip code data cell in our sample included 12 residents – the data miner could possibly find out who the customers are. Having the exact geographical location of a customer would be desirable to determine user profiles more accurately. However, this would infer privacy problems because precise coordinates could reveal a customer's identity.

Inferences in our analysis framework depend on the secondary data that is available for

[20] Often referred to as "reidentification" or "triangulation" problems

data linkage. Voter registration lists containing attributes such as `name`, `zip_code` and `date_of_birth` that were used by Sweeney [2001] to reidentify hospital patients are not publicly available in Germany because federal law prohibits access of third parties to ones voter registry (§17 I of the German Federal Electoral Law [BWahlG, 2005]). Thus, individuals in Germany are likely to be less affected by inference problems than those described in Sweeney [2001] due to limited access to external information.

In addition to identified inference problems, there is an inherent risk that future inference problems may impact a company's analysis framework.

4.2.4 Problem statement

In summary, certain purposes are allowed in the analysis of certain data, and the data may be used for this purpose by certain recipients. The basic relational framework of P3P, however, is insufficient to account for inferences that may *substitute* certain data. Regardless of whether their use was permitted or not, data are *available* or not, and for each indicator, certain data are *required*. Legal regulations may *restrict* data usage. These relations constitute the problem specification for the analysis prototype (cf. Figure 4-2).

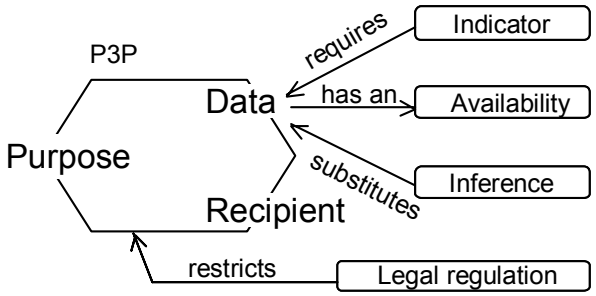

Figure 4-2: Problem specification

4.3 Design

This section presents the prototype design. Section 4.3.1 presents the main data types and relations used in the prototype. Section 4.3.2 describes the main functions and work processes. An extension of P3P for the privacy requirements presented in Section 4.2 is proposed.

4.3.1 Data types and relations

We distinguish between data the analyses computation is working on (input data) and data the process is working with (process data). The input data is formed by the Web log and Web user data (such as purchase, socio-economic, geographic and other data), together with the privacy policy the enterprise has adopted. Physically, this policy consists of the P3P file. The process data is the business logic that defines the whole analysis

process.

4.3.1.1 Input data

The input data describe the data items, purposes, recipients in Figure 4-2 as well as the relations between them. The input data consists of three sets: the set of basic data elements *D*, the set of purposes *P* and the set of recipients *R*. These are the same entities as those defined in P3P. Note that all these sets are enumerable:

D = {user, third party, business, dynamic} with every element again a set of data, as it is defined in Cranor et al. [2002]. Note that *D* can be extended by the issuer of the policy. Furthermore, we define for further use D_{set} which is formed by sets of elements of *D*.

P is the set of the 12 relevant purposes as defined for the PURPOSE element, and *R* is the set of the six possible values for the RECIPIENT element. These two sets are not extensible.

The P3P STATEMENT establishes a relation between elements belonging to these three groups by assembling the DATA, PURPOSE and RECIPIENT elements.

4.3.1.2 Process data

The analysis framework introduces two new data entities for the process data. The first is the set of analyses *I* which is formed by all metrics and analytics that can be calculated from the present data. This set is fixed and not user-extensible. The second entity is the availability *A = {true, false}*. *A* indicates whether an instance of data is physically stored in the enterprise and can be made available to the analysis process. Note that this availability is defined purely technically. No privacy aspects are considered at this point.

4.3.1.3 Functional data relations

The functional data describe the relations between the availability, the analyses and the data items in Figure 4-2. Before analyzing the functional relations between the different data, we introduce our notation of functional relationship [Pepper, 2003]. A function *f* is a triple $(\mathfrak{D}_f, \mathfrak{W}_f, \mathfrak{R}_f)$, formed by a domain $\mathfrak{D}_f$, a range $\mathfrak{W}_f$ and a relation $\mathfrak{R}_f$, the function graph. This function graph has to be injective, i.e. there are no two pairs $(a, b_1) \in \mathfrak{R}_f$ and $(a, b_2) \in \mathfrak{R}_f$ with $b_1 \neq b_2$. The function f maps the argument value *x* to the result value *y* if the pair (*x*, *y*) is part of the function graph: $(x, y) \in \mathfrak{R}_f$. A given function $f = (\mathfrak{D}_f, \mathfrak{W}_f, \mathfrak{R}_f)$ is called partial if

$\pi_1(\mathfrak{R}_f) \subset \mathfrak{D}_f$, where π_1 is the projection defined as $\pi_1(A \times B) = A$[21]. Otherwise, i.e. if $\pi_1(\mathfrak{R}_f) = \mathfrak{D}_f$, *f* is called total.

Every statement in a given policy is an implicit function definition of a function *h* as: $h: D \times R \times P \rightarrow \{allowed\}$. The codomain of this implicit function is the one-element set *{allowed}*. This function is (usually) partial as not all purposes are allowed to everyone for all the data. In the following, we will totalize *h* by defining *k* as $k(\underline{x}) = h(\underline{x})$ if $\underline{x} \in \mathfrak{R}_h$ and $k(\underline{x}) =$ *{not allowed}* otherwise. *k* is total.

Example: consider a statement as a fragment of a P3P file such as the following excerpt from Example 4.1 in Cranor et al. [2002]:

```
 ...
<STATEMENT>
     <PURPOSE><individual-decision
     required="optout"/></PURPOSE>
     <RECIPIENT><ours/></RECIPIENT>
     RETENTION><stated-purpose/></RETENTION>
     <DATA-GROUP>
          <DATA ref="#user.name.given"/>
          <DATA ref="#dynamic.cookies">...</DATA>
     </DATA-GROUP>
</STATEMENT>
  ...
```

This fragment defines the following elements of $\mathfrak{R}_h$:

((user.name.given, ours, individual-decision), allowed)

((dynamic.cookies, ours, individual-decision), allowed)

There are two more functions that establish relations:

The function *requiredfor*: $D \times I \rightarrow \{true, false\}$ defined on the data *D* and the analyses *I* states whether a data item is used within the calculation of an analysis. The function

[21] Analogously: $\pi_2(A \times B) = B$

isavailable: $D \rightarrow A$ indicates whether a given data item is available. By definition, *isavailable(<>) = true* where <> indicates "no data".

As all the sets are enumerable, these functions *k*, *requiredfor* and *isavailable* can be defined "point for point" for all elements. They are deterministic. Extensions of *D* require an extension of all three function graphs.

4.3.2 Functions and work processes

Given the set of all possible analyses, the subset "executable analyses" is the set of all the analyses that can be executed. We define each of our metrics and analytics in the analysis framework as a business analysis *I*. So, $I \supseteq I_{executables} = t(I)$ where the function *t* selects all the executable analyses from *I* (*t* acts as a filter). This section will provide the definition of *t*: $I \rightarrow I$.

Whether a given analysis $i \in I$ is executable or not depends on two requirements: its execution has to be feasible and its execution must be allowed. With respect to an implementation of the framework, it is reasonable to check in this order because the check for technical requirements is usually simpler.

The technical requirements are (i) the presence of the definition of this analysis (the implementation has to know how to calculate it) – we presume that this is always guaranteed – and (ii) the presence of the data that is needed for its calculation, i.e. $isavailable(d_j)=true\ \forall d_j$: $requiredfor(d_j, i)=true$.

The restrictions imposed by the privacy policy are expressed by *k*. The execution of an analysis *i* is allowed if $k(d_j, r, p)=allowed\ \forall d_j$: $requiredfor(d_j, i)=true$ where $r \in R$ and $p \in P$ have to be specified by the analyst.

There is no fixed relation between purpose and analysis, as the calculation of a given analysis can have multiple purposes. As discussed in Section 4.1, the lack of such relations is a serious problem for the privacy-compliant analysis of consumer data if one party is mistrusted. For each analysis result the prototype displays the P3P purpose as a relation of data attribute and specified purpose. For analyses combining data items with different usage specifications an alert message is also displayed.

We define *t*, the filter for the executable analyses, as a composition of functions already known (<> is "no analysis"):

$$t(i)=\begin{cases} i & \text{if } (isavailable(d_j)=true) \\ & \wedge\ k(d_j, r, p)=allowed \\ & \nexists d_j: requiredfor(d_j, i)=true \\ <> & \text{otherwise} \end{cases}$$

In this consideration, we have assumed that a company stores its data with attribute names and level of aggregation as defined by the P3P base data schema. In real systems, this assumption is usually not fulfilled. Additional matching and aggregation or disaggregation of data have to be done. But as this is only a question of naming and storing, it has no impact on the theoretical process of decision making.

4.3.2.1 Impact of data inference on decision making

We define an inference as a function s: $D_{set} \rightarrow D$. If there is an inference, then we can write $s(\{d_1, d_2, \ldots, d_n\}) = d_{n+1}$ with $d_i \neq d_j \Leftrightarrow i \neq j$. The existence of inferences is a problem for the decision on whether an analysis can be calculated or not. In particular, there is a problem for $\Re_h$.

Consider two data items for which the same restrictions on purpose and recipient apply: (d_1, r_1, p_1) and (d_2, r_1, p_1). Moreover, there is an inference so that $s(\{d_1,d_2\})=d_3$. For d_3, the following purpose limitation applies: (d_3, r_1, p_3). Consider an analysis that the recipient r_1 wants to use for the purpose p_1 which requires the data d_3. Calculating this analysis by d_3 directly is prohibited by the P3P policy if the desired purpose is different from the allowed purpose ($p_1 \neq p_3$). However, calculating the analysis from d_1 and d_2 is possible. Thus, inferences may bypass privacy restrictions.

The site user who accepted the policy is not protected against this violation of her privacy preferences – unless she employs a user agent that (i) is aware of this inference possibility and (ii) extends the usage restriction to also cover inferred data.

To achieve this goal, we propose an extension to P3P. Additional elements can be included into a policy by the element EXTENSION as defined in Cranor et al. [2002].

We suggest an unordered list of inference statements. Each INFERENCE statement consists of the data that can be inferred if a given set of data is present. A human-readable explanation can be added within the CONSEQUENCE element.

From a given premise, it may be possible to conclude n consequences. This is expressed as n separate INFERENCE statements, each with an atomic consequence. In addition, one may want to express an inference possibility such as “if d_1 and (d_2 or d_3) are given, then it is possible to infer d_4”. This may be split into two statements: “if d_1 and d_2, then d_4”

and “if d_1 and d_3, then d_4”. However, the introduction of the connector OR in addition to AND makes the formulation and reading of inferences easier for human users.

DATA-GROUPs can be placed within one of these elements to express logical relations between them. The following fragment shows an example.

```
...
<EXTENSION optional="no">


<INFERENCES xmlns="http://cleo.wiwi.hu-
 berlin.de/simt/extensions">


  <INFERENCE>
    <CONSEQUENCE> If the zip code and the birth
     date are known, the home address can be
     reconstructed. </CONSEQUENCE>

    <GIVEN>

      <AND>

        <DATA-GROUP>
          <DATA ref="#user.home-info.
           postal.country"/>
          <DATA ref="#user.home-info.
           postal.stateprov"/>
          <DATA ref="#user.home-info.
           postal.postalcode"/>
          <DATA ref="#user.bdate"/>

        </DATA-GROUP>
      </AND>
    </GIVEN>


    <INDUCED>

      <DATA-GROUP>
        <DATA ref="#user.home-info.
         postal.street"/>
      </DATA-GROUP>
    </INDUCED>

  </INFERENCE>


  <INFERENCE>
```

```
    <CONSEQUENCE> The international telephone
     code can be reconstructed from the name of
     the country, and vice versa.</CONSEQUENCE>
    <GIVEN>
      <OR>
        <DATA-GROUP>
          <DATA ref="#user.home-info.
           postal.country"/>
        </DATA-GROUP>
        <DATA-GROUP>
          <DATA ref="#user.home-info.
           telecom.telephone.intcode"/>
        </DATA-GROUP>
      </OR>
    </GIVEN>

    <INDUCED>
      <DATA-GROUP>
        <DATA ref="#user.home-info.
         postal.country"/>
        <DATA ref="#user.home-info.
         telecom.telephone.intcode"/>
      </DATA-GROUP>

    </INDUCED>

  </INFERENCE>


</INFERENCES>


</EXTENSION>

...
```

User agents should parse these inferences. As this extension adds further restrictions to the policy, it is mandatory.

According to the W3C specification of P3P we define an INFERENCES extension using the Augmented Backus-Naur Form (ABNF) notation of [Crocker and Overel]. For simplicity, we abstain from an XML schema definition, even though a loss of flexibility has to be taken into account.

```
inferences = "<INFERENCES>" 1*inference "</INFERENCES>"

inference =  "<INFERENCE>"
```

```
                consequence
                given
                induced
                "</INFERENCE>"

given =         "<GIVEN>" logical "</GIVEN>"


induced =       "<INDUCED>" data-group "</INDUCED>"


logical =       or_set | and_set


or_set =        "<OR>" ((1*data-group)|logical)
                "</OR>"
and_set =       "<AND>" ((1*data-group) | logical)
                "</AND>"
```

4.3.2.2 Coding legal restrictions in a P3P policy

As we have pointed out in Section 4.2.1, laws impose restrictions on using data. These restrictions are usually independent of recipient and purpose [EU, 2002]. Whereas the STATEMENTs in a policy file allow using the data within the specified borders, legal specifications always restrict uses. A priori, one can say that any legal restriction can be coded in a P3P policy by listing all allowed uses. Thus, the missing uses are prohibited. But this realization does not respect the simultaneity restriction: consider two data d_1 and d_2 that can be used by a given recipient r_1 for a given purpose p_1. These separate uses are allowed by the laws and so may be listed in a P3P policy. But combining (i.e. simultaneous use) the same data for the same purpose is not allowed. This restriction cannot be coded by a P3P policy. Thus we suggest the introduction of a new element LEGAL that restricts combined usage in order to remedy this lack of P3P.

Within the LEGAL element several RESTRICTION elements can be specified. Each RESTRICTION can have four attributes; the introduction of additional attributes or values needs to be discussed. The ISSUER attribute specifies the name of the legal authority that codified the restriction, the LAW attribute contains the name (possibly shortened) of the legal norm which is the origin for this restriction. The values of both attributes are human-readable strings. The FOR-attribute indicates the region the site user must belong

to for this restriction to be applied. Possible values are comma-separated combinations of “all”, “EU”, and the ISO country abbreviations such as “US” for the United States of America, “GB” for the United Kingdom, or “DE” for Germany[22]. The default value is “all”. Finally, the non-value attribute “viceversa” summarizes the repetition of the same restriction with reversed WHILE and DON’T elements:

```
<RESTRICTION viceversa>
     <WHILE> A  </WHILE>
     <DONT>  B  </DONT>
</RESTRICTION>
```

is equivalent to:

```
<RESTRICTION>
     <WHILE> A  </WHILE>
     <DONT>  B  </DONT>
</RESTRICTION>
<RESTRICTION>
     <WHILE> B  </WHILE>
     <DONT>  A  </DONT>
</RESTRICTION>
```

Within the RESTRICTION element, a CONSEQUENCE element can be defined, as it is defined in the P3P specification and also used for the extension by INFERENCE.

The main elements are WHILE and DONT. Both of them contain a single DATA-GROUP with one or more DATA elements. The use of all the DATA in the DONT element concurrently with the DATA in the WHILE element is not allowed. As this extension adds further restrictions to the policy that cannot be ignored, it is a mandatory extension.

The following fragment shows an example of the P3P extension using the LEGAL-element.

```
<LEGAL>
  <RESTRICTION
```

[22] http://www.iso.org

```
   issuer="European Commission"
   law="EU Privacy Directive"
   for="EU"
   viceversa>

    <CONSEQUENCE> Information about site
     usage is not allowed to be combined with
     identifiable personal user data.
    </CONSEQUENCE>


    <WHILE>
      <DATA-GROUP>
        <DATA ref="#user.name"/>
      </DATA-GROUP>
    </WHILE>

    <DONT>

      <DATA-GROUP>

        <DATA ref="#dynamic.clickstream"/>
        <DATA ref="#dynamic.http"/>
      </DATA-GROUP>

    </DONT>

  </RESTRICTION>

</LEGAL>
```

As the same legal restrictions apply for a large variety of Web sites, mechanisms to include a set of referenced legal restrictions hosted by a trusted provider (e.g. governmental authorities) should be developed as well.

4.3.2.3 Workflow

Figure 4-3 summarizes the processes within the framework, including the successive data exchanges and actions between the involved participants. The analysis provider's task "identifies inferences, executes / disables analyses" is both an action and a restriction. For each exchange, its format is noted in an exemplary form. Interunit exchanges rely on standardized protocols and data description formats. Note that the framework includes the extensions for legal restrictions and inference problems.

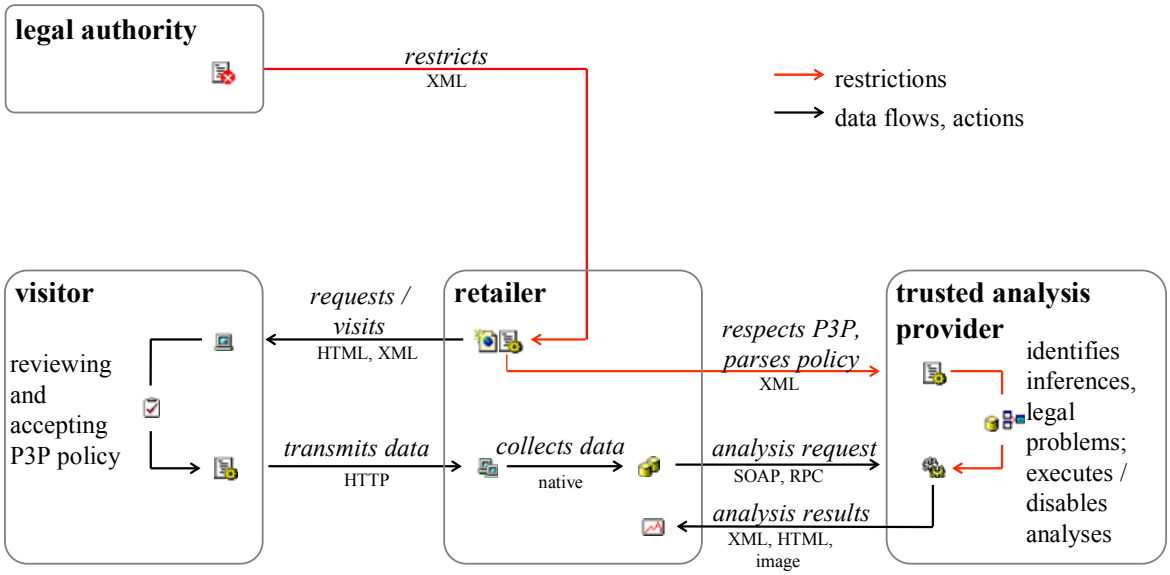

Figure 4-3: Workflow

4.4 User interface

We have implemented a prototype based on the analysis framework proposed in Chapter 3. This section is reserved for a (non-complete) technical description of the prototype.

The analysis service has three specification phases. Currently, the specification has to be done manually. Future releases will support automated data retrieval and policy parsing. In each of the three phases, the analyst is told her specific task. Input errors are directly reported.

The first phase is the specification of the data the enterprise has stored: data availability is defined here. The second phase is the specification of the P3P privacy policy that applies to the data specified in the first step. The third phase is the selection of the analysis time frame and the desired analyses. The list of 82 metrics and analytics grouped in eight categories is presented. The user interface only shows the metrics and analytics that are allowed given available data and legal privacy restrictions. Other analyses are disabled and displayed in grayish color. The time frame (time interval of analysis) can be typed directly or chosen from a calendar control. Once an analysis has been chosen, a set of three output formats are proposed depending on the type of analysis: output as HTML, as XML or as an image. Images are generated dynamically using standard classes of the .NET Framework. The analyst can handle this image like all other images – she can save it, copy it, etc. Image formats (PNG, GIF, JPEG, BMP, TIFF, etc.), colors and fonts can be freely configured. The direct streaming avoids problems with asynchronous page request, image generation and image request. Moreover, there are no problems with temporary files. During our analyses based on the data of the multi-channel retailer, no time lags

were detected. The image generation “on the fly” does not slow down the output flush.

Figure 4-4 shows a screen shot of the analysis tool user interface (phase 3 of the specification process). In the background you can see a part of the analyses choice list with some choices disabled:

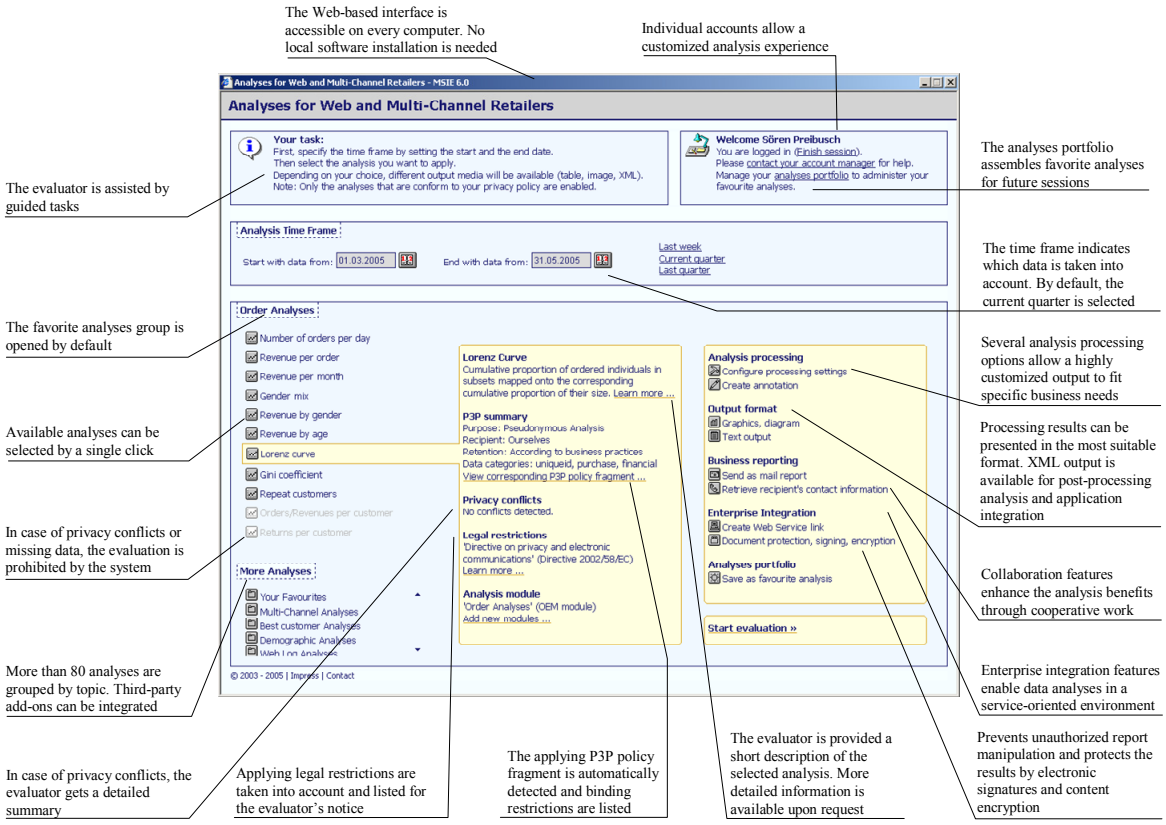

Figure 4-4: Main interface design with analyses choice list, privacy indication and time frame selection

4.5 Implementation

The prototype is a Web-based application written in C# in Microsoft .Net. The Web server dynamically generates Web pages to interact with the analyst who is not required to install additional client software. All browser types are supported as long as they support clientside ECMAScript (JScript or JavaScript).

Two databases are involved in the analysis process: the first is a Microsoft (MS) Access database providing the complete preprocessed Web data to be analyzed. The second is a MS SQL Server database that holds the process data.

According to the P3P Guiding Principles [Cranor, et al., 2002], measures have been taken to implement mechanisms for protecting any information that is transferred from the analyst to the tool and vice versa. HTTP over a high SSL encryption is used as a trusted

protocol for the secure transmission of data. Restrictive session timeouts prevent the abuse of foreign sessions. Analysts have to log on with a personal password, and temporary session cookies are used to prevent other analysts from “stealing” a session.

The system has been tested on data from the described multi-channel retailer. The application of the service framework on an online retailer’s consumer data indicated privacy problems in a real-world context. Potential inference problems and legislative privacy implications were identified and could be addressed within the framework.

4.6 Modification of analyses

In the hypothetical case that the analysis service is untrustworthy and the data collector wishes to protect the data before transfer, we briefly discuss possible protection measures. One solution for protecting sensitive data in a two-party business case – as described in Section 4.1 – is encryption techniques. The basic idea is to leave the data on the data collector’s server and transfer only encrypted data to the service provider [cf. Domingo-Ferrer and Herrera-Joancomarti, 1999; Rivest, et al., 1978]. Asonov and Freytag [2002] described a hardware-based approach to encryption using a secure coprocessor. Encryption functions are useful for a limited number of algorithmic operations such as addition, subtraction, multiplication and inverse multiplication and for basic database queries such as selection, projection and join [Boyens, 2004]. However, for more complex mining queries such as those in our analysis framework, encryption techniques are not suited.

Statistical disclosure control is an approach to minimizing privacy problems in databases [Agrawal and Srikant, 2000; Willenborg and Waal, 2001]. Statistical disclosure control can be broadly classified into the groups of query restriction and data perturbation [Agrawal and Srikant, 2000]. Using these techniques, data will be modified in such a way that the probability of reidentifying individual users can be kept below a selected threshold. Query restriction includes the restriction of the size of query results [Denning, et al., 1979; Fellegi, 1972], the control of overlap amongst successive queries [Dobkin, et al., 1979], the suppression of data cells of small size [Cox, 1980], and the clustering of entities into mutually exclusive atomic populations [Yu and Chin, 1977]. Perturbation techniques suggest ways of adding noise to the data while maintaining some statistical invariant. Perturbation techniques include the swapping of values between records (Denning 1982), the replacement of the original data by a sample from the same distribution [Lefons, et al., 1983], the adding of noise to the results of a query [Beck, 1980], and the sampling of query results [Denning, 1982]. Both methods have advantages and disadvantages and none is an optimal solution: query restriction cannot completely avoid inferences but

provides valid responses. Perturbation techniques can prevent inferences but may not provide precise query results.

For our analysis framework, the following disclosure techniques are particularly useful for minimizing privacy problems:

1. *Limit access to the data*, i.e. hide attributes that potentially identify data subjects (e.g. `customer_id, address, email,` exclude zip code areas with a small number of inhabitants)
2. *Aggregate the data*, i.e. summarize the data in such a way that no conclusions can be drawn for a single subject, e.g. use zip codes instead of more fine-grained location data or replace the exact `date_of_birth` with the `year_of_birth`.
3. *Assign unique identifiers randomly*, i.e. deploy primary keys that do not contain additional information about the subject they are pointing to, e.g. do not assign `customer_id` in consecutive order because it could possibly be linked with a person-related IP number.

A problem with these disclosure techniques could be a limited quality of the query results. For example, in the case of geomarketing, the shop obviously needs to make a trade-off between the preciseness of its results and the potential privacy violation of its users.

Moreover, inference opportunities could pose a privacy risk. Inference problems that are not known at the time of anonymization could inherently threaten user privacy in statistical disclosure control [Boyens, 2004].

4.7 Conclusion

A framework for deploying Web analyses has been set up and tested on data from a multi-channel retailer. We have determined the different data types that are involved in the data analysis process and established the functional relations between them. An automated way of filtering business analyses according to privacy restrictions has been presented. Due to our proposed extensions of the P3P specification, it is now possible to code both data inferences and legal usage restrictions.

We proposed approaches to modify the analyses in such a way that they could be transferred to an untrusted service provider.

“Not everything that can be counted counts, and not everything that counts can be counted.” (Albert Einstein)

5 Extension of user privacy requirements

Chapter 4 discussed the impact of privacy restrictions specified in legal frameworks and P3P policies on our analysis framework presented in Chapter 3. As indicated in Chapter 3 the results from the framework can be particularly useful for Web site personalization. As personalization systems become more effective with an increasing amount of user information, the impact of consumer privacy concerns is particularly high for these applications. This chapter discusses privacy concerns from a consumer point of view in more detail. We will compare 30 consumer privacy surveys, categorize them and point out the particular implications for personalization systems.

This chapter is organized as follows: Section 5.1 defines characteristics of personalization. Section 5.2 categorizes personalization systems according to the input data they require. Section 5.3 summarizes privacy concerns from more than 30 consumer surveys and describes their impact on personalization systems. Differences between consumers’ privacy views and their actual behaviors, and differences between consumer and industry opinions on privacy are also presented. Section 5.4 discusses future research directions and proposes approaches on how to increase consumer trust in personalization systems.

5.1 User-adaptable vs. user-adaptive systems

Personalized (or “user-adaptive”) systems have gained substantial momentum with the rise of the WWW. The market research firm Jupiter [Foster, 2000] defines personalization as predictive analysis of consumer data used to adapt targeted media, advertising and merchandising to consumer needs. A more Web-oriented definition was proposed by [Kobsa, et al., 2001] who regard a personalized hypermedia application as a hypermedia system that adapts the content, structure and/or presentation of the networked hypermedia objects to each individual user’s characteristics, usage behavior and/or usage environment. In contrast to user-adaptable systems where the user is in control of the initiation, proposal, selection and production of the adaptation, user-adaptive systems perform all steps autonomously.

The advantages of personalization can be manifold. Web site visitors see the major benefits in sites being able to offer more relevant content and to recall user preferences and interests [Cyber Dialogue, 2001]. The personalization of hypermedia is beneficial for several other purposes as well, most notably for improving the learning progress in

educational software [Brusilovsky, et al., 1998; Specht, 1998]. Given the increasing amount of information offered on the Internet, the development of advanced personalized services seems to become inevitable.

Personalization systems need to acquire a certain amount of data about users' interests, behavior, demographics and actions before they can start adapting to them. Thus, they are often useful in domains only where users engage in extended (and most often repeated) system use. They may not be appropriate for infrequent users with typically short sessions. The extensive and repeated collection of detailed user data, however, may provoke consumer privacy concerns. Consumer surveys show that the number of consumers refusing to shop online because of privacy concerns is as high as 64% [Culnan and Milne, 2001]. Finding the right balance between privacy protection and personalization remains a challenging task.

5.2 Input data for personalization

Kobsa [2001] divides the data that are relevant for personalization purposes into 'user data', 'usage data', and 'environment data'. 'User data' denote information about personal characteristics of a user, while 'usage data' relate to a user's (interactive) behavior (e.g. as captured in the Web log). A special kind of 'usage data' is 'usage regularities', which describe frequently reoccurring interactions of users. 'Environment data' refer to the user's software and hardware, and the characteristics of the user's current locale.

Table 5-1 lists the most frequently occurring subtypes of these data. The taxonomy allows one to refer to specific kinds of personalization systems more easily, and facilitates our analysis of privacy concerns and their impacts on certain system types.

No.	Input Data	Examples of User-Adaptive Systems
A) User Data:		
I	Demographic Data	Personalized Web sites based on user profiles; software providers: Broadvision, Personify, NetPerceptions etc.
II	User Knowledge	Expertise-dependent personalization; product and technical descriptions: Sales Assistant [Popp and Lödel, 1996], SETA [Ardissono and Goy, 2000]; learning systems: KN-AHS [Kobsa, et al., 1994], [Brusilovsky, 2001]

III	User Skills and Capabilities	Help Systems: Unix Consultant [Chin, 1989], [Küpper and Kobsa, 1999]; disabilities: AVANTI [Fink, et al., 1998]
IV	User Interests and Preferences	Recommender systems [Resnick and Varian, 1997]; used car domain: [Jameson, et al., 1995]; domain of telephony devices: [Ardissono and Goy, 1999]
V	User Goals and Plans	Personalized support for users with targeted browsing behavior, plan recognition: [Lesh, et al., 1999], PUSH [Höök, et al., 1996], HYPERFLEX [Kaplan, et al., 1993]

B) Usage Data:

VI	Selective Actions	Adaptation based on link-selection: WebWatcher [Joachims, et al., 1997], Letizia [Lieberman, 1995]; image-selection: Adaptive Graphics Analyser [Holynski, 1988]
VII	Temporal Viewing Behavior	Adaptation based on viewing time; streaming objects: [Joerding, 1999]; temporal navigation behavior: [Chittaro and Ranon, 2000]; micro-interaction: [Sakagami, et al., 1998]
VIII	Ratings	Adaptation based on object ratings; product suggestions: Firefly [Shardanand and Maes, 1995], GroupLens [Konstan, et al., 1997]; Web pages: [Pazzani and Billsus, 1997]
IX	Purchases and Purchase-related actions	Suggestions of similar goods after product selection: Amazon.com; other purchase-related actions: registering, transferring products into virtual shopping cart, quizzes
X	Other (dis-) confirmatory actions	Adaptation based on other user actions, e.g. saving, printing documents, bookmarking a Web page: [Konstan, et al., 1997]

C) Usage Regularities:

XI	Usage Frequency	Adaptation based on usage frequency; icon toolbar: [Debevc, et al., 1996], Flexcel [Krogsaeter, et al., 1994]; Web page visits: AVANTI [Fink, et al., 1998]
XII	Situation-action correlations	Interface agents; routing mails: [Mitchell, et al., 1994], [Maes, 1994], meeting requests: [Kozierok and Maes, 1993]
XIII	Action Sequences	Recommendations based on frequently used action sequences, e.g. past actions, action sequences of other users

D) Environment Data:

XIV	Software Environment	Adaptation based on users' browser versions and platforms, availability of plug-ins, Java and JavaScript versions
XV	Hardware Environment	Adaptation based on users' bandwidth, processor speed, display devices (e.g. resolution), input devices
XVI	Locale	Adaptation based on users' current location (e.g. country code), characteristics of usage locale

Table 5-1: Types of personalization-relevant data and examined systems

5.3 Results from privacy surveys

5.3.1 Impacts on user-adaptive systems

We categorized 30 recent consumer surveys on Internet privacy (or summaries of such surveys), and analyzed their potential impacts on the different types of personalization systems listed in Table 5-1 (summary of taxonomy in Kobsa et al. [2001]). Questions from different surveys addressing the same privacy aspects were grouped together, to convey a more complete picture of user concerns. 11 documents included all questions, six provided an extensive discussion of survey results, and 10 contained factual executive summaries. For three studies, only press releases were available.

We distinguished several categories of privacy aspects. The category 'privacy of user data in general' has a direct impact on any personalization system that requires personal data

(such as the user's name, address, income etc.). The category 'privacy in a commercial context' primarily affects personalized systems in e-commerce. 'Tracking of user sessions' and 'use of cookies' influence user-adaptive systems requiring usage data. A few studies focus on 'e-mail privacy'. This category might have an impact on user-adaptive systems that generate targeted e-mails. Two studies directly address the topic of privacy and personalization [Mabley, 2000; Personalization Consortium, 2000]. They are highly interesting because they directly affect most personalization systems.

Results regarding user data in general	Systems affected
Internet users who are concerned about the security of personal information: 83% [Cyber Dialogue, 2001], 70% [Behrens, 2001], 72% [UMR, 2001], 84% [Fox, et al., 2000]	I, II, IV, V, IX
People who have refused to give (personal) information to a Web site: 82% [Culnan and Milne, 2001]	I, II, IV, V, IX
Internet users who would never provide personal information to a Web site: 27% [Fox, et al., 2000]	I, II, IV, V, IX
Internet users who supplied false or fictitious information to a Web site when asked to register: 34% [Culnan and Milne, 2001], 24% [Fox, et al., 2000]	I, II, IV, V, IX
Online users who think that sites who share personal information with other sites invade privacy: 49% [Cyber Dialogue, 2001]	I, II, IV, V, IX, XIII

Table 5-2: Results regarding user data in general

A significant concern about the use of personal information can be seen in these results, which is a problem for those personalization systems in Table 5-1 that require 'user data' (such as demographic data', data about 'user knowledge', etc.). Systems that record 'purchases and purchase-related actions' may also be affected. More than a quarter of the respondents even indicated that they would never consider providing personal information to a Web site. Quite a few users indicated having supplied false or fictitious information to a Web site when asked to register, which makes user linking across sessions and thereby accurate recommendations based on 'user interests and preferences' very difficult.

Results regarding user data in a commercial context	Systems affected
People wanting businesses to seek permission before using their personal information for marketing: 90% [Roy Morgan Research, 2001]	I, II, IV, V, IX
Non-online shoppers who did not purchase online because of privacy concerns: 66% [Ipsos Reid, 2001], 68% [Interactive Policy, 2002], 64% [Culnan and Milne, 2001]	I, II, IV, V, IX
Online shoppers who would buy more if they were not worried about privacy/security issues: 37% [Forrester, 2001], 20% [Department for Trade and Industry, 2001]	I, II, IV, V, IX
Shoppers who abandoned online shopping carts because of privacy reasons: 27% [Cyber Dialogue, 2001]	I, II, IV, V, IX
People who are concerned if a business shares their data for a different than the original purpose: 91% [UMR, 2001], 90% [Roy Morgan Research, 2001]	IX, XIII

Table 5-3: Results regarding user data in a commercial context

These results suggest that in a commercial context, privacy concerns may play an even more important role than for general personalized systems. Most people want to be asked before their personal information is used, and many regard privacy as a must for Internet shopping. Thus, commercial personalization systems need to include privacy features. In particular, those systems in Table 5-1 that require ‘demographic data’, ‘user knowledge’, ‘user interests and preferences’, ‘user goals and plans’ and ‘purchase-related actions’ are affected.

Furthermore, more than 90% of respondents are concerned if a business shares their information for a different than the original purpose. This has a severe impact on central user modeling servers that collect data from, and share them with, different user-adaptive applications, unless sharing can be controlled by the user [Kobsa, 2001; Kobsa and Schreck, 2003].

Results regarding user tracking and cookies	Systems affected
People who are concerned about being tracked on the Internet: 60% [Cyber Dialogue, 2001], 54% [Fox, et al., 2000], 63% [Harris Interactive, 2000]	VI-X, XIV-XVI
People who are concerned that someone might know what Web sites they visited: 31% [Fox, et al., 2000]	VI-X, XIV-XVI
Internet users who generally accept cookies: 62% [Personalization Consortium, 2000]	VI-X, XIV-XVI
Internet users who set their computers to reject cookies: 25% [Culnan and Milne, 2001], 3% [Cyber Dialogue, 2001], 31% in warning modus [Cyber Dialogue, 2001], 10% [Fox, et al., 2000]	VI-X, XIV-XVI
Internet users who delete cookies periodically: 52% [Personalization Consortium, 2000]	VI-X, XIV-XVI
Users uncomfortable with schemes that merge tracking of browsing habits with an individual's identity: 82% [Harris Interactive, 2000]	I, II, IV-X, XIV-XVI
User who feel uncomfortable being tracked across multiple Web sites: 91% [Harris Interactive, 2000]	VI-X, XIV-XVI, XIII

Table 5-4: Results regarding user tracking and cookies

Users' privacy concerns about tracking and cookies affect the acceptance of personalization systems based on 'usage data' and 'usage regularities' (cf. Table 5-1). In particular, systems using 'selective actions', 'temporal viewing behavior' and 'action sequences' conflict with users' privacy preferences. More than 50% of Internet users are concerned about Internet tracking [Cyber Dialogue, 2001; Fox, et al., 2000]. Fox et al. [2000] found that user tracking is not welcome even when users receive personalized content in return. A significant number claimed they would set their browser to reject cookies [Culnan and Milne, 2001; Mabley, 2000] and more than half of the users stated they would delete cookies periodically [Personalization Consortium, 2000].

The results directly affect machine-learning methods that operate on user log data since without cookies, sessions of the same user cannot be linked any more. User concerns of tracking schemes across multiple Web sites affects personalization systems that combine

information from several sources, in particular those systems that use data from 'action sequences', 'demographics', 'purchase-related actions' and the user's 'locale'.

Most users do not consider current forms of tracking as helpful methods to collect data for personalization. Users' participation in deciding when and what usage information should be tracked might decrease such privacy concerns.

Results regarding e-mail privacy	Systems affected
People who have asked for removal from e-mail lists: 78% [Cyber Dialogue, 2001], 80% [Culnan and Milne, 2001]	XII
People who complain about irrelevant e-mail: 62% [Ipsos Reid, 2001]	XII
People who have received unsolicited e-mail: 95% [Cyber Dialogue, 2001]	XII
People who have received offensive e-mail: 28% [Fox, et al., 2000]	XII

Table 5-5: Results regarding e-mail privacy

In the category of e-mail privacy, 62% of the users complain about irrelevant e-mail [Ipsos Reid, 2001]. Almost every Internet user has already received unsolicited e-mail [Mabley, 2000]. This may constitute a problem for the acceptance of personalized e-mail. The problem affects primarily those systems in Table 5-1 that use 'situation-action correlation'. The findings indicate that many deployed e-mail personalization systems, such as software for the management of targeted marketing campaigns, are not yet able to address user needs specifically enough to evoke positive reactions among the recipients.

Results regarding privacy and personalization	Systems affected
Online users who see personalization as a good thing: 59% [Harris Interactive, 2000]	I-XVI
Online users who do not see personalization as a good thing: 37% [Harris Interactive, 2000]	I-XVI
Types of information users are willing to provide in return for personalized content: name: 88%, education: 88%, age: 86%, hobbies: 83%, salary 59%, credit card number: 13% [Cyber Dialogue, 2001]	I, II, IV, V, IX

Internet users who think tracking allows the site to provide information tailored to specific users: 27% [Fox, et al., 2000]	VI-X, XIV-XVI
Online users who think that sites who share information with other sites try to better interact: 28% [Cyber Dialogue, 2001]	I-XVI
Online users who find it useful if a site remembers information (preferred colors, delivery options etc.): 50% [Personalization Consortium, 2000]	I-V, XIV-XVI, IX
People who are bothered if a Web site asks for information one has already provided (e.g., mailing address): 62% [Personalization Consortium, 2000]	I-V, XIV-XVI, IX
People who are willing to give information to receive a personalized online experience: 51% [Personalization Consortium, 2000], 40% [Roy Morgan Research, 2001], 51% [Privacy & American Business, 1999]	I-V, IX

Table 5-6: Results regarding privacy and personalization

The results of the study by Harris Interactive [2000] affect all systems in Table 5-1. A significant portion of the respondents does not seem to see enough value in personalization that they would be willing to give out personal data. If any possible, personalization should therefore be designed as an option that can be switched off. Finally, Internet users also demonstrated less commitment to providing personal information in return for personalized content when a Web site would share this information with other sites. This result applies to all personalized systems that share information via a central user modeling server [Kobsa, 2001].

5.3.2 Differences in consumer statements and actual privacy practices

This meta-analysis demonstrates that consumers are highly concerned about the privacy implications of various data collection methods, but many would share some data in return for personalization.[23] Users however do not seem to always have a good understanding of

[23] Users' willingness to share information with a Web site may also depend on other factors that are not considered here such as the usability of a site, users' general level of trust towards a site, and the company or industry to which the site belongs. For example, good company reputation makes 74% of the surveyed Internet users more comfortable disclosing personal information [Ipsos Reid, 2001].

their privacy needs in a personalization context. Stated privacy preferences and actual behavior often diverge:

- User tracking evokes significant privacy concerns, but only 10% (27%) of American Internet users have set their browsers to reject cookies [Fox, et al., 2000; Roy Morgan Research, 2001].
- 76% of survey respondents claimed that privacy policies on Web sites were very important to them [Behrens, 2001], but in fact users barely view such pages when visiting Web sites [Kohavi, 2001].
- In an experiment, [Berendt, et al., 2005] found that users often do not live up to their self-reported privacy preferences: subjects claimed to be highly concerned about their privacy, but shared very personal and sensitive information with a personalized Web site.

5.3.3 Differences in the privacy views of consumers and industry

Besides differences in consumers' self-perception and actual behavior, our analysis of survey results also uncovered a few major discrepancies in the privacy views of consumers and industry. Consumer expectations and actual industry practices should however be in line with each other, so that consumers can build trust which is the basis for the acceptance of personalization. For instance, 54% do not believe that most businesses handle the personal information they collect in a proper and confidential way [Harris Interactive, 2003; Responsys.com, 2000]. In contrast, 90% of industry respondents believe that this is the case for their own business, and 46% that this is the case for industry in general.[24]

Consumer demands and current practice in companies also diverge significantly on the issue of data control. Most Internet users (86%) believe that they should be allowed control over what information is stored by a business [Fox, et al., 2000], but only 17% of businesses allow users to delete at least some personal information [Andersen Legal, 2001]. Furthermore, 40% of businesses do not provide access to personal data for verification, correction and updates [Deloitte Touche Tohmatsu, 2001].

Industry and consumers also disagree significantly on the value of privacy laws. Nine of

[24] However, only 40% of businesses say steps have been taken to secure personal information held by a site (Internet Privacy Survey 2001), and 55% do not store personal data in encrypted form. 15% share user data with third parties without having obtained users' permission (Deloitte 2001).

ten marketers claim that the current regime of self-regulation works for their companies, and 64% think that government involvement will ultimately hurt the growth of e-commerce [Responsys.com, 2000]. In contrast, two-thirds of e-mail users think that the federal government should pass more laws to ensure citizens' privacy online [Gallup Organization, 2001], while only 15% supported self-regulation [Harris Interactive, 2000]. However, it has been found that trust in the effectiveness of privacy legislation has meanwhile decreased among consumers [Harris Interactive, 2001].

Although both governments and private organizations have made serious efforts to ease users' privacy concerns, much remains to be done to build and maintain customer confidence, which is a prerequisite for successful personalization.

5.3.4 Discussion of the methodology

The cited studies were mostly conducted by well-known research institutions and market research firms between 2000 and 2003. The number of respondents in the studies varied between 500 and 4500, with an average of about 2000. The answers were collected by telephone interviews and online questionnaires. From the 30 surveys analyzed, 21 were conducted in the US, three in Canada, two in Australia and New Zealand, two in Britain and one in the European Union. One survey was based on an international respondent sample.

Though this meta-analysis provides a more comprehensive and objective overview of privacy concerns and their impacts on personalization than can be expected from a single study, some caution should be exercised. A general problem is the lack of comparability of the studies: small differences in the wording of the questions, their context in the questionnaires, the recruitment method and the sample population make user statements difficult to compare. Harper and Singleton [2001] criticized the use of manipulative questions in many privacy studies, a lack of trade-offs between privacy and other desires, and imprecise terminology (e.g. the term "privacy" is often understood as a synonym for security, or a panacea against identity fraud and spam). Finally, as mentioned above, disparities seem to exist between people's responses to general, context-less privacy questions, and their behavior when working with concrete Web sites having specific goals in mind.

5.4 Conclusion

Our meta-analysis of consumer surveys demonstrated that users' privacy concerns are major. Survey results regarding Web user data in general, Web user data in a commercial context, Web usage data, e-mail privacy and personalization have been discussed. The impact of privacy concerns on personalization systems has been described.

Two different directions can be pursued to alleviate these concerns. In one approach, users receive commitments that their personal data will be used for specific purposes only, including personalization. Such commitments can be given in, e.g., individual negotiations or publicly displayed privacy promises ("privacy policies"), or they can be mandated in privacy laws as discussed in Section 4.2.1. It is necessary though that these privacy commitments be guaranteed. They ought to be enforced through technical means [Agrawal, et al., 2002; Fischer-Hübner, 2001; Karjoth, et al., 2003], or otherwise through audits and legal recourse. Since individual privacy preferences may considerably vary between users, Kobsa [2003] proposes a meta-architecture for personalized systems that allows them to cater to individual privacy preferences and to the privacy laws that apply to the current usage situation. The personalized system would then exhibit the maximum degree of personalization that is permissible under these constraints.

The other approach is to allow users to remain anonymous with regard to the personalized system and the whole network infrastructure, whilst enabling the system to still recognize the same user in different sessions so that it can cater to her individually [Kobsa and Schreck, 2003]. Karat, Brodie, Karat, Vergo and Alpert [2003] also address this requirement through different levels of identity. Anonymous interaction seems to be desired by users (however, only a single user poll addressed this question explicitly so far [GVU, 1998]). One can expect that anonymity will encourage users to be more open when interacting with a personalized system, thus facilitating and improving the adaptation to the respective user. As discussed in Section 4.2.1, the anonymous use of data can relieve the providers of personalized systems from restrictions and duties imposed by such laws (they may however choose to observe these laws nevertheless, or to provide other privacy guarantees on top of anonymous access).

It is currently unclear which of these two directions should be preferably pursued. Each alternative has several advantages and disadvantages. Neither is a full substitute for the other, and neither is guaranteed to alleviate users' privacy concerns, which ultimately result from a lack of trust. For the time being, both directions need to be pursued.

6 Contextualized communication of privacy practices and personalization benefits

The meta-study of consumer privacy surveys in Chapter 5 demonstrated that today's Web users are becoming increasingly privacy-conscious and less willing to disclose personal data to companies. Thus, respecting privacy requirements in data usage as described in Chapter 4 is not enough. A Web site must also effectively communicate these privacy practices to its users in order to increase trust and willingness to buy online respectively.

As discussed in Chapter 5, privacy protection is particularly important in user-adaptive Web sites, as they require more detailed user information than regular sites and therefore pose higher privacy risks. Thus, Web sites need more advanced methods for communicating to users both their privacy practices and the benefits that users can expect by providing personal data. In this chapter, we will discuss and analyze such methods in the context of personalized Web sites [Kobsa, et al., 2001].

More than two-thirds of the respondents in Ackerman et al. [1999] indicated that knowing how their data will be used would be an important factor in their decision on whether or not to disclose personal data. It seems, though, that the communication of privacy practices on the Internet has so far not been very effective in alleviating consumer concerns: 64% of Internet users surveyed in Culnan and Milne [2001] indicated having decided in the past not to use a Web site, or not to purchase something from a Web site, because they were not sure about how their personal information would be used.

Currently, the predominant way for Web sites to communicate how they handle users' data is to post comprehensive privacy statements. 76% of users find privacy policies very important [Department for Trade and Industry, 2001], and 55% stated that a privacy policy makes them more comfortable disclosing personal information [GartnerG2, 2001; Roy Morgan Research, 2001]. However, privacy statements today are usually written in a form that gives the impression that they are not really supposed to be read. And this is indeed not the case: whereas 73% of the respondents in Harris Interactive [2000] indicate having viewed Web privacy statements in the past (and 26% of them claim to always read them), Web site operators report that users hardly pay any attention to them [Kohavi, 2001]. Abrams [2003] criticizes that people are turned off by long, legalistic privacy notices whose complexity makes them wonder what the organization is hiding. We clearly need better means for communicating corporate privacy practices than what is afforded by today's privacy statements on the Web.

Communicating a company's privacy policy alone is not sufficient though. In situated

interviews [Brodie, et al., 2004], users pointed out that “in order to trust an e-commerce company, they must feel that the company is doing more than just protecting their data – it must also be providing them with functionality and service that they value”. The way in which personal data is used for the provision of these services must be clearly explained. Current Web privacy statements hardly address the connection between personal data and user benefits.

The following section will survey existing approaches to communicating privacy practices to Web site visitors that go beyond the posting of privacy statements, and indicate their merits and shortcomings. Section 6.2 proposes a new contextualized strategy to communicate privacy practices and personalization benefits. Section 6.3 presents a new interface design approach for a sample Web site. In Section 6.4, we describe a between-subjects experiment in which we compare this approach with a traditional form of disclosure. We focus on differences between users’ willingness to share personal data, differences in their purchase behavior, and differences in their perception of a site’s privacy practices as well as the benefits they received by sharing their data. Section 6.5 concludes this chapter with a discussion of results.

6.1 Existing approaches and their shortcomings

As discussed in Section 4.2.2, the currently predominant approach to communicating privacy practices to Web site visitors besides privacy statements is the Privacy Preferences Protocol (P3P). However, the current P3P adoption rate is stagnating [Ernst&Young, May 2004]. One reason may be due to P3P’s problematic legal implications (as discussed in Section 4.2.2) and the insufficient support to users in evaluating a Web site’s P3P policy.

The latter problem is partly addressed by the AT&T Privacy Bird [AT&T, 2002], which allows users to specify their own privacy preferences, compares them with a site’s P3P-encoded privacy policy when users visit this site, and alerts them when this policy does not meet their standards. Upon request, the Privacy Bird also provides a summary of a site’s privacy policy and a statement-by-statement comparison with the user’s privacy preferences.

A few browsers also allow users to specify certain limited privacy preferences and to compare them with the P3P policies of visited Web sites. For example, Internet Explorer 6 allows users to initially state a few privacy preferences and blocks cookies from sites that do not adhere to these preferences. The Mozilla browser goes one step further and allows users to enter privacy settings for cookies, images, popup windows, certificates and smart cards.

Finally, a simple non-technical approach is suggested by [Abrams, 2001; Abrams, 2003]. The author correctly points out that the current lengthy and legalistic privacy statements "don't work". As an alternative, he suggests a "layered approach" which includes: one short concise notice with standardized vocabulary that is easy to follow and highlights the important information, and an additional long, "complete" policy that includes the details.

All these approaches suffer from the following major shortcomings though:

1. They require users to make privacy decisions upfront, without regard to specific circumstances in the context of a particular site or of individual pages at a site. This disregards the situational nature of privacy [Palen and Dourish, 2002]. In fact, privacy preferences stated upfront and actual usage behavior often seem to differ significantly [Berendt, et al., 2005; Spiekermann, et al., 2001].
2. The systems do not inform about the benefits of providing the requested data. For instance, respondents in (Personalization Consortium, 2000) indicate to be willing to share personal data if the site offered personalized services.
3. They do not enhance users' understanding of basic privacy settings. For example, most users still do not know what a cookie is and what it can do.

Very recent work takes first steps to address some of these deficiencies. Friedman, Howe and Felten [2002] aim at further enhancing the above-mentioned management of cookies and users' privacy in the Mozilla browser. Among other things, the authors study contextual issues such as how to enhance users' understanding of cookie settings, at the time when cookie-related events occur and in a form that is least distractive. Patrick and Kenny [2003] is concerned with the communication of privacy choices under the European Data Protection Directive [EU, 2002]. From the privacy principles of this Directive, the authors derive four HCI guidelines for effective privacy interface design: (1) comprehension, (2) consciousness, (3) control, and (4) consent. Since single large click-through privacy policies or agreements do not meet the spirit of the Directive, the authors propose "just-in-time click-through agreements" on an as-needed basis instead of a large, complete list of service terms. These small agreements would facilitate a better understanding of decisions since they are made in-context.

6.2 A communication design pattern

To adequately address privacy concerns of users of personalized Web sites, we propose user interface design patterns that communicate the privacy practices of a site both at a global and a local level. Similar to design patterns in object-oriented programming, interface design patterns constitute descriptions of best practices within a given design domain based on research and application experience [van Duyne, et al., 2002]. They

give designers guidelines for the efficient and effective design of user interfaces.

6.2.1 Global communication

Global communication of privacy practices currently takes place by posting privacy statements on a company's home page or on all its Web pages. As pointed out in Section 4.2 privacy statements on the Web are legally binding in many jurisdictions. Privacy policies are therefore carefully crafted by legal council. Rather than completely replacing them by something new whose legal impact is currently unclear at best, our approach keeps current privacy statements in the "background" for legal reference and protection. However, we argue to enhance this kind of disclosure by additional information that explains privacy practices and user benefits, and their relation to the requested personal data, in the given local context.

6.2.2 Local communication

As discussed in Section 6.1, tailored in-context explanation of privacy practices and personalization benefits can be expected to address users' privacy concerns much better than global contextless disclosures. Such an approach would break long privacy policies into smaller, more understandable pieces, refer more concretely to the current context, and thereby allow users to make situated decisions regarding the disclosure of their personal data considering the explicated privacy practices and the explicated personalization benefits.

It is unclear yet at what level of granularity the current context should be taken into account. Should privacy practices and personalization benefits be explained at the level of single entry fields (at the risk of being redundant), or summarized at the page level or even the level of several consecutive pages (e.g., a page sequence for entering shipping, billing and payment data)? Several considerations need to be taken into account:

Closure: Input sequences should be designed in such a way that their completion leads to (cognitive) closure [Shneiderman and Plaisant, 2004]. The coarsest level at which closure should be achieved is the page level. This therefore should also be the coarsest level for the provision of information about privacy and personalization, even if this information is redundant across several pages.

Separation: Within a page, sub-contexts often exist that are supposed to be visually separated from each other (e.g. simply by white space). Ideally, the completion of each sub-context should lead to closure. Information about privacy and personalization should therefore be given at the level of such visually separated sub-contexts, even if this leads to redundancy across different contexts on a page.

Different sensitivity: Ackerman et al. [1999] found that users indicated different degrees of willingness to give out personal data, depending on the type of data and whether the data was about them or their children. For instance, 76% of the respondents felt comfortable giving out their own email addresses, 54% their full names, but only 11% their phone numbers. Even when entry fields for such data fall into the same sub-context (which is likely in the case of this example), users' different comfort levels suggest to treat each data field separately and to provide separate explanations of privacy practices and personalization benefits that can address these different sensitivity levels.

Legal differences: From a legal perspective, not all data may be alike. For instance, the European Data Protection Directive distinguishes "sensitive data" (such as race, ethnic origin, religious beliefs and trade union membership) whose processing require the user's explicit consent. This calls for a separate explanation of privacy practices and personalization benefits of data that are different from a legal standpoint, possibly combined with a "just-in-time click-through agreement" as proposed by Patrick and Kenny [2003].

The safest strategy is seemingly to communicate privacy practices and personalization benefits at the level of each individual entry field for personal data. If a number of such fields form a visually separate sub-context on a page, compiled explanations may be given only if the explanations for each individual field are not very different (due to legal differences, different sensitivity levels, privacy practices or personalization benefits). A page is the highest possible level at which compiled contextual explanations may be given (again, only if the field-level explanations are relatively similar). Visually separate sub-contexts on a page should be preferred though, due to the closure that they require.

6.3 Interface design pattern of an example Web site

Figure 6-1 shows the application of the proposed interface design pattern to a Web bookstore that offers personalized services. The top three links in the left-hand frame lead to the global disclosures (to facilitate comprehension, we decided to split the usual contents of current privacy statements into three separate topics: privacy, personalization benefits, and security). The main frame contains input fields and checkboxes for entering personal data. Each of them is accompanied by an explanation of the site's privacy practices regarding the respective personal data (which focuses specifically on usage purposes), and the personalized services that these data afford.

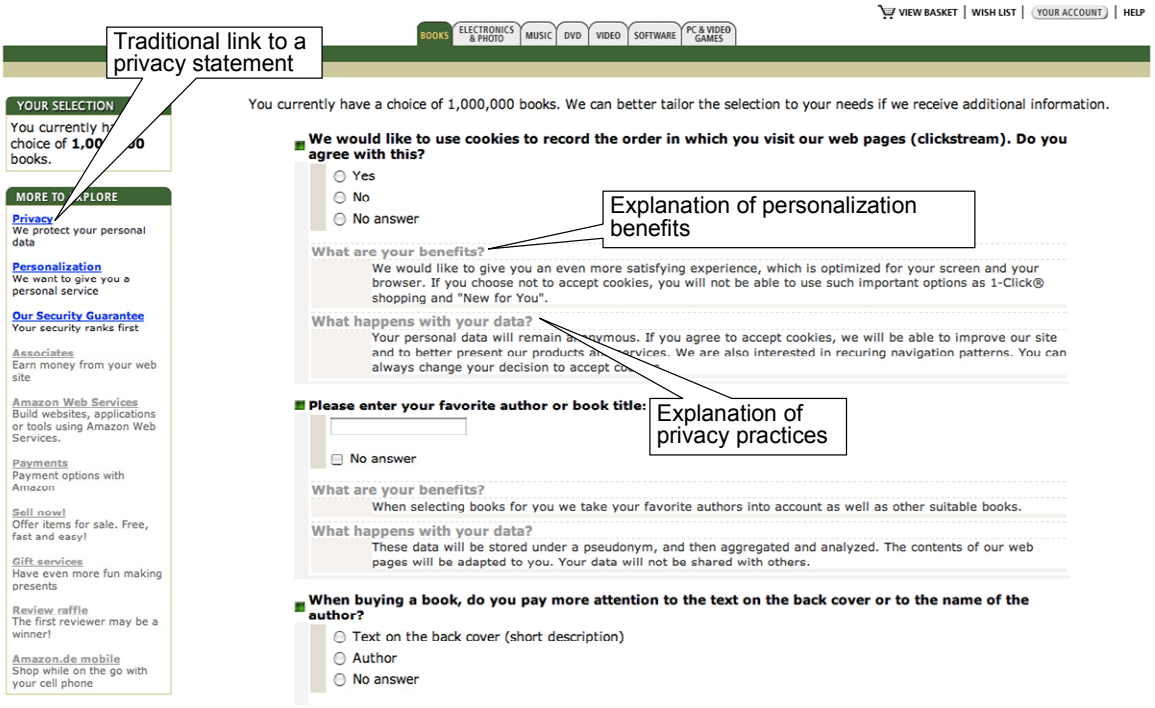

Figure 6-1: Global and contextual communication of privacy practices and personalization benefits

As in the theoretical model of Lederer, Dey and Mankoff [2002], a user achieves an understanding of the privacy implications of the displayed situation both intuitively (taking the overall purpose of the site and page into account) and through adequate contextual notice. The traditional link to a privacy policy can still be accessed if so desired.

6.4 Impacts on users' data sharing behavior

We conducted a user experiment to empirically verify the merits of our proposed user interface design pattern in comparison with traditional approaches for the communication of privacy practices. In Section 6.4.1 we will motivate the specific research strategy that we pursued. Sections 6.4.2-6.4.5 describe the materials, subjects, design and procedures, and the results of our study.

6.4.1 Background

Two kinds of methods can be applied to study users' reaction to different interface designs: inquiry-based and observational methods. In the first approach, users are being interviewed about their opinions with regard to the questions at hand. These interviews may be supported by representations of the proposed designs, ranging in fidelity from paper sketches to prototypes and real systems. In the second approach, users are being observed while carrying out tasks (either their customary ones or synthetic tasks). Both approaches complement each other: while inquiries may reveal aspects of users' rationale

that cannot be inferred from mere observation, observations allow one to see actual user behavior which may differ from self-reported behavior.

This latter problem seems to prevail in the area of privacy. Berendt et al. [2005] and Spiekermann et al. [2001] found that users' stated privacy preferences deviate significantly from their actual behavior, and an enormous discrepancy can be observed between the number of people who claim to read privacy policies and the actual access statistics of these pages. Solely relying on interview-based techniques for analyzing privacy impacts on users, as is currently nearly exclusively the case, must therefore be viewed with caution. Our empirical studies therefore gravitated towards an observational approach, which we complemented by questionnaires. We designed an experiment to determine whether users exhibit different data sharing behavior depending on the type of explanation about privacy practices and personalization benefits that they receive (global alone versus global plus contextual). Our hypothesis was that users would be more willing to share personal data in the condition with contextual explanations, and that they would also view sites more favorably that use this type of disclosure.

6.4.2 Materials

We developed a fake book recommendation and sales Web site whose interface was designed to suggest an experimental future version of a popular online bookstore. Two variants of this system were created, one with contextual explanations of privacy practices and personalization benefits, and one without. Figure 6-1 shows an excerpt of the first variant, translated from German into English. The contextual explanations are given for each entry field (which is the safest of the strategies discussed in Section 6.2.2), under the headings "What are your benefits?" and "What happens with your data?" In the version without contextual explanations, these explanations are omitted.

In both conditions, the standard privacy policy of the Web retailer is used. The three left-hand links labeled "Privacy", "Personalization" and "Our Security Guarantee" lead to the original company privacy statement (we split it into these three topics though and left out irrelevant text). In the condition with contextual explanations, the central policies that are relevant in the current situation are explained under "What happens with your data?" Such explanations state, for instance, that the respective piece of personal data will not be shared with third parties, or that some personal data will be stored under a pseudonym and then aggregated and analyzed. The explanation of the usage purpose is concise and kept in the spirit of P3P specifications [Cranor, et al., 2002].

A counter was visibly placed on each page that purported to represent the size of the currently available selection of books. Initially the counter is set to one million books. Data

entries in Web forms (both via checkboxes and radio buttons and through textual input) decrease the counter after each page by an amount that depends on the data entries made. The Web forms ask a broad range of questions relating to users' interests. A few sensitive questions on users' political interests, religious interests and adherence, their literary sexual preferences, and their interest in certain medical subareas (including venereal diseases) are also present. All questions "make sense" in the context of filtering books in which users may be interested. For each question, users have the option of checking a "no answer" box or simply leaving the question unanswered. The personal information that is solicited in the Web forms was chosen in such a way that it may be relevant for book recommendations and/or general customer and market analysis. Questions without any clear relation to the business goals of an online bookstore are not being asked. A total of 32 questions with 86 answer options are presented. Ten questions allow multiple answers, and seven questions have several answer fields with open text entries (each of which we counted as one answer option). The complete set of questions and their contextual explanations are provided in the Appendix to Chapter 6.

After nine pages of data entry (with a decreased book selection count after each page), users are encouraged to review their entries and then to retrieve books that purportedly match their interests. 50 predetermined and invariant books are then displayed that were selected based on their low price and their presumable attractiveness for students (book topics include popular fiction, politics, tourism, and sex and health advisories). The prices of all books are visibly marked down by 70%, resulting in out-of-pocket expenses between 2 and 12 euros for a book purchase. For each book, users can retrieve a page with bibliographic data, editorial reviews, and ratings and reviews by readers.

Users are free to choose whether or not to buy one single book. Those who do are asked for their shipping and payment data (a choice of bank account withdrawal and credit card charge is offered). Those who do not buy may still register with their postal and email addresses, to receive personalized recommendations in the future as well as newsletters and other information.

6.4.3 Subjects

58 subjects participated in the experiment. They were students of Humboldt University in Berlin, mostly in the areas of Business Administration and Economics. The data of six subjects were eventually not used, due to a computer failure or familiarity with the student experimenters.

6.4.4 Experimental design and procedures

The experiment was announced electronically in the School of Economic Sciences of

Humboldt University. Participants were promised a 6 euros coupon for a nearby popular coffee shop as a compensation for their participation, and the option to purchase a book with a 70% discount. Prospective participants were asked to bring their IDs and credit or bank cards to the experiment.

When subjects showed up for the experiment, they were reminded to check whether they had these credentials with them, but no data was registered at this time. Paraphernalia that are easily associated with the Web book retailer, such as book cartons and logos, were casually displayed.

In the instructions part of the experiment, subjects were informed that they would test an experimental new version of the online bookstore with an intelligent book recommendation engine inside. Users were told that the more and the better data they provided, the better would be the book selection. They were made aware that their data would be given to the book retailer after the experiment. It was explicitly pointed out though that they were not required to answer any question. Subjects were asked to work with the prototype to find books that suited their interests, and to optionally pick and purchase one of them at a 70% discount. They were instructed that payments could be made by credit card or by withdrawal from their bank accounts. The student briefing is provided in the Appendix of Chapter 6.

A between-subjects design was used for the subsequent experiment, with the system version as the independent variable: one variant featured non-contextual explanations of privacy practices and personalization benefits only, and the other additionally contextualized explanations (see Section 4.2 for details). Subjects were randomly assigned to one of the two conditions (we will abbreviate them by “no-ctxt-expl” and “ctxt-expl” in the following). They were separated by screens, to bar any communication between them. After searching for books and possibly buying one, subjects filled in a post-questionnaire on paper. The questionnaire is provided in the Appendix of Chapter 6. Finally, the data of those users who had bought a book or had registered with the system were compared with the credentials that subjects had brought with.

6.4.5 Results

Data Sharing Behavior. We analyzed the data of 26 participants in the conditions “no-ctxt-expl” and “ctxt-expl”. We first dichotomized their responses by counting whether a question received at least one answer or was not answered at all. Whereas on average 84% of the questions were answered in condition “no-ctxt-expl”, this rose to 91% in the second condition (see Table 6-1). A Chi-Square test on a contingency table with the total number of questions answered and not answered in each condition showed that the

difference between conditions was statistically significant ($p < 0.001$).

The two conditions also differed with respect to the number of answers given (see Table 6-2). The maximum number of answers that any subject could give was 64, and we used this as the maximum number of possible answers. In condition “no-ctxt-expl”, subjects gave 56% of all possible responses on average (counting all options for multiple answers), while they gave 67% of all possible answers in condition “no-ctxt-expl”. A Chi-Square contingency test showed again that the difference between the two conditions is highly significant ($p < 0.001$). The relative difference between the number of answers provided in the two conditions is even higher than in the dichotomized case (19.6% vs. 8.3% increase).

	w/o contextual explanations	with contextual explanations	df	Chi-Square	*p*	N
% Questions answered	84%	91%	1	16.42	< 0.001	1664

Table 6-1: Percentage of questions answered and results of Chi-Square test

	w/o contextual explanations	with contextual explanations	df	Chi-Square	*p*	N
% Answers given	56%	67%	1	42.68	< 0.001	3432

Table 6-2: Percentage of checked answer options and results of Chi-Square test

The results demonstrate that the contextual communication of privacy practices and personalization benefits has a significant positive effect on users’ willingness to share personal data. The effect is even stronger when users can give multiple answers. We found no evidence for a significant difference of this effect between questions that we regarded as more sensitive and less sensitive questions.

Purchases. Table 6-3 shows that the purchase rate in condition “ctxt-expl” is 33% higher than in condition “no-ctxt-expl” (note that all subjects saw the same set of 50 books in both conditions). A t-test for proportions indicates that this result approaches significance ($p < 0.07$). We regard this as an important confirmation of the success of our proposed contextual explanation of privacy practices and personalization benefits. In terms of privacy, the decision to buy is a significant step since at this point users reveal personally identifiable information (name, shipment and payment data) and risk that previously pseudonymous information may be linked to their identities. A contextual explanation of

privacy practices seemingly alleviates such concerns much better than a traditional global disclosure of privacy practices.

	w/o contextual explanations	with contextual explanations	df	t	*p(t) (1-tailed)*	N
Purchase ratio	0.58	0.77	48	1.51	0.07	52

Table 6-3: Purchase ratio and result of t-test for frequencies

Access to the global company disclosures. We also monitored how often subjects clicked on the links “Privacy”, “Personalization” and “Our Security Guarantee” in the left side panel (which lead to the respective original global company disclosures): merely one subject in each condition clicked on the “Privacy” link.

Rating of privacy practice and perceived benefit resulting from data disclosure. The paper questionnaire that was administered to each subject at the end of the study contains five Likert questions (whose possible answers range from “strongly agree” to “strongly disagree”), and one open question for optional comments. It examines how users perceive the level of privacy protection at the Web site as well as the expediency of their data disclosure in helping the company recommend better books.

The responses to the five attitudinal questions were encoded on a one to five scale. A one-tailed t-test revealed that the agreement with the statement “Privacy has priority at <book retailer>” was significantly higher in condition “ctxt-expl” than in condition “no-ctxt-expl” ($p < 0.01$). The same applies to subjects’ perception of whether their data disclosure helped the bookstore in selecting interesting books for them ($p < 0.05$). Note again that all subjects were offered the same set of books. The difference between the two conditions in the statement “<book retailer> uses my data in a responsible manner” approached significance ($p < 0.12$). More details about these results can be found in Table 6-4.

Item	N	no-ctxt-expl Means	no-ctxt-expl Std Dev	ctxt-expl Means	ctxt-expl Std Dev	$Means_{dif}$	$Std\ Dev_{dif}$	t	df	*p(t) 1-tailed*
Privacy has priority	41	3.35	0.88	3.94	0.87	0.60	0.28	2.16	39	0.01
Data helped site to select better books	56	2.85	0.97	3.40	1.10	0.51	0.28	1.85	54	.035
Data is used responsibly	47	3.62	0.85	3.91	0.83	0.29	0.25	1.17	45	0.12

Table 6-4: Users' perception of privacy practice and benefit of data disclosure

6.5 Discussion and open research questions

Our experiment was designed so as to ensure that subjects had as much "skin in the game" as possible, and thereby to increase its ecological relevance. The incentive of a highly discounted book and the extremely large selection set that visibly decreased with every answer given was chosen to incite users to provide ample and truthful data about their interests. The perceptible presence of the Web book retailer, the claim that all data would be made available to them, and the fact that names, addresses and payment data were verified (which ensured that users could not use escape strategies such as sending books to Post Office boxes or someone they know) meant that users really had to trust the privacy policy that the Web site promised when deciding to disclose their identities.

The results demonstrate that the contextualized communication of privacy practices and personalization benefits has a significant positive effect on users' data sharing behavior, and on their perception of the Web site's privacy practices as well as the perceived benefit resulting from data disclosure. The additional finding that this form of explanation also leads to more purchases approached significance. The adoption by Web retailers of interface design patterns that contain such explanations therefore seems clearly advisable.

While the experiment does not allow for substantiated conclusions regarding the underlying reasons that link the two conditions with the observed effects, the results are by all means consistent with recent models in the area of personalization research that include the notion of 'trust' in a company (cf. Chapter 2). One may speculate whether the significantly higher perceived usefulness of data disclosure in condition "ctxt-expl" can be explained by a positive transfer effect.

Other characteristics of our experiment are also in agreement with the literature. Hine and Eve [1998] found in their study of consumer privacy concerns that "in the absence of straightforward explanations on the purposes of data collection, people were able to

produce their own versions of the organization's motivation that were unlikely to be favorable. Clear and readily available explanations might alleviate some of the unfavorable speculation" [emphasis ours]. Culnan and Bies [2003] postulate that consumers will "continue to disclose personal information as long as they perceive that they receive benefits that exceed the current or future risks of disclosure. Implied here is an expectation that organizations not only need to offer benefits that consumers find attractive, but they also need to be open and honest about their information practices so that consumers […] can make an informed choice about whether or not to disclose." The readily available explanations of both privacy practices and personalization benefits in our experiment meet the requirements spelled out in the above quotations, and the predicted effects could be indeed observed.

Regarding our results, we would like to point out that additional factors may also play a role in users' data disclosure behavior, which were kept constant in our experiment due to the specific choice of the Web retailer, its privacy policy, and a specific instantiation of our proposed interface design pattern. We will discuss some of these factors in the following.

Reputation of a Web site. We chose a Web store that enjoys a relatively high reputation in Germany (we conducted surveys that confirmed this). It is well known that reputation increases users' willingness to share personal data with a Web site [cf. CG&I-R, 2001; Earp and Baumer, 2003; Teo, et al., 2004]. Our high response rates of 84% without and specifically 91% with contextual explanation suggest that we may have already experienced some ceiling effects (after all, some questions may have been completely irrelevant for the interests of some users so that they had no reason to answer them). An experiment with a retailer who has a lower perceived reputation should be conducted.

Stringency of a Web site's data handling practices. The privacy policy of the Web site that we mimicked is comparatively strict. Putting this policy upfront and explaining it in-context in a comprehensible manner is more likely to have a positive effect on customers than couching it in legalese and hiding it behind a link. Chances are that this may change if a site's privacy policy is not so customer-friendly.

Permanent visibility of contextual explanations. In our experiment, the contextual explanations were permanently visible. This uses up a considerable amount of screen real estate. Can the same effect be achieved in a less space-consuming manner, for instance with icons that symbolize the availability of such explanations? If so, how can the contextual explanations be presented so that users can easily access them and at the same time will not be distracted by them? Should this be done through regular page links, links to pop-up windows, or rollover windows that pop up when users brush over an icon?

References to the full privacy policy. Privacy statements on the Web currently constitute important and comprehensive legal documents. Contextual explanations will in most cases be incomplete since they need to be short and focused on the current situation, so as to ensure that users will read and understand them. For legal protection, it is advisable to include in every contextual explanation a proviso such as “This is only a summary explanation. See <link to privacy statement> for a full disclosure.” Will users then be concerned that a Web site is hiding the juicy part of its privacy disclosure in the “small print”, and therefore show less willingness to disclose their personal data? Additional user experiments will be necessary to obtain answers or at least a clearer picture with regard to these questions.

7 Conclusion and future research

The objective of this thesis was to propose solutions for mitigating potential conflicts of interests regarding online privacy and data use between companies and customers. A particular emphasis was placed on the business model of multi-channel retailing that dominates e-commerce.

In Chapter 2 it was shown that cross-channel effects exist between a company's physical store network and its e-shop. The *perceived size* and *reputation* of physical stores had a significant influence on consumers' *perceived trust* in the e-shop. *Perceived privacy* had the most important influence on the development of consumer *trust* in our model. The results motivated our further research on privacy and multi-channel retailing.

Chapter 3 developed a Web analysis framework with 82 analyses for Web sites. New conversion success metrics and customer segmentation approaches have been proposed. A particular emphasis has been placed on the development of metrics and analytics for multi-channel retailers. The metrics have been calculated for a data sample of Web user and usage data from a European multi-channel retailer and an information Web site. Implications of the results have been discussed and recommendations for improving business success online have been derived.

Chapter 4 integrated privacy requirements into the analysis process. A privacy-preserving Web analysis tool has been developed for the analyses defined in Chapter 3. The tool indicates when business analyses are not compliant with legal privacy regulations or P3P specifications and thus supports the privacy management within a company. A syntactical extension of P3P for known inference problems and legal regulations has been proposed.

Chapter 5 provided an overview of consumer privacy concerns. A meta-study of 30 privacy surveys emphasized the importance of consumer privacy regarding Web user and usage data. Moreover, the impact of privacy concerns on user-adaptive systems has been discussed. Possible solutions to privacy-preserving personalization have been suggested.

Chapter 6 proposed a new user interface design approach, in which the privacy practices of a Web site were explicated in a contextualized manner and users' benefits in providing personal data clearly explained. A user experiment has been conducted that compared two versions of a personalized Web store: one with a traditional global disclosure and one that additionally provides contextualized explanations of privacy practices and personalization benefits. Subjects in the second condition were significantly more willing to share personal data with the Web site, rated its privacy practices and the perceived benefit resulting from data disclosure significantly higher, and also made considerably

more purchases.

Regarding the structural equation model in Chapter 2 future work should further focus on the interactive influence between e-shop and physical stores. It could be that a reciprocal effect from the Internet on physical stores exists. A comparison of the mutual effects should further explain why multi-channel retailing is such a successful business strategy. Moreover, the integration of further “media channels” (mail, television) and “institutional channels” (call center, sales force) would be an interesting research aspect. Further work should also analyze the impact of cultural differences on privacy perceptions [cf. Jarvenpaa, 1999].

The Web analysis framework in Chapter 3 proposed five categories of Web analyses with 82 metrics and analytics. The framework could be enhanced by integrating further metrics and analytics. In particular, cost-related and detailed product-related analyses would be a useful extension.

The analysis framework was tested on data from a retailer who sells consumer electronics, which belong to a product category that is successfully sold on the Internet [Omwando, 2002]. A discussion of the impact of product characteristics such as search and experience attributes on Internet suitability has been discussed in related work [cf. Nelson, 1974; Phau and Poon, 2000; Subramaniam, et al., 2000; Wright and Lynch, 1995]. Further work should discuss the impact of product characteristics on the analysis results in more detail. For example, in Teltzrow et al. [2003a] we have shown that consumers tend to increasingly pick up products with increasing product weight and price.

In Chapter 4, a further development of the privacy-preserving analysis prototype is envisioned. The Web service can be improved by codifying different privacy policies that can be extended to customers from which they can choose one according to their desired privacy level. From a legal viewpoint, it would be interesting to develop a set of Web analyses that meet the requirements of privacy regulations in different countries. A matching and combination of service providers’ policies, user-defined P3P preferences and legal restrictions would help to better protect the customer’s informational self-determination.

An extension of the business model could be the exchange of analysis results between several companies using the framework. Further privacy questions would arise that need to be solved in the context of a three-party business case [Boyens, 2004]. Developing encryption techniques for an analysis Web service that guarantees anonymized transfer of data back and forth from a company to a service provider is also a research question that should be addressed in further work.

In Chapter 5 we discussed the impact of privacy on personalization. Further research should focus on solutions to privacy-preserving personalization [Kobsa, 2003]. Experimental studies should further analyze what company commitments and/or technical solutions increase consumer trust in online personalization.

Regarding our privacy communication design in Chapter 6, additional factors should be tested that may also play a role in users' data disclosure behavior. In particular, further experiments should explore whether the reputation of a Web site, the stringency of a Web site's data handling practices, the visibility of contextual explanations or the placing of references to the full privacy policy have an impact on data disclosure and willingness to buy.

References

Abrams, M. (2001): The Notices Project: Common Short Informing Notices, Interagency Public Workshop: Get Noticed: Effective Financial Privacy Notices, Washington, DC. URL: http://www.ftc.gov/bcp/workshops/glb/presentations/abrams.pdf

Abrams, M. (2003): Making Notices Work For Real People., Proceedings of the 25th International Conference of Data Protection & Privacy Commissioners, Sydney, Australia. URL: http://www.privacyconference2003.org/

Ackerman, M.S.; Cranor, L. and Reagle, J. (1999): Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences, Proceedings of the 1st ACM E-Commerce Conference, Denver, Co.

Agrawal, R.; Bayardo, R.J.; Faloutsos, C.; Kiernan, J.; Rantzau, R. and Srikant, R. (2004): Auditing Compliance with a Hippocratic Database, Proceedings of the 30th International Conference on Very Large Data Bases, Toronto, Canada.

Agrawal, R.; Imielinski, T. and Swami, A. (1993): Mining Associations between Sets of Items in Massive Databases, Proceedings of the ACM-SIGMOD International Conference on Management of Data, Washington D.C.

Agrawal, R.; Kiernan, J.; Srikant, R. and Xu, Y. (2002): Hippocratic Databases, 28th International Conference on Very Large Databases, Hong Kong, China. URL: http://www.vldb.org/conf/2002/S05P02.pdf

Agrawal, R. and Srikant, R. (2000): Privacy-Preserving Data Mining, ACM-SIGMOD 2000 Conference on Management of Data, Dallas, TX.

Agre, P.E. and Rotenberg, M. (1997): Technology and Privacy: The New Landscape, MIT Press, Cambridge, MA.

Ajzen, I. (1991): The theory of planned behavior, Organizational Behavior and Human Decision Processes (vol. 50), pp. 179-211.

Alba, J.; Lynch, J.; Weitz, B.; Janiszewski, C.; Lutz, R.; Sawyer, A. and Wood, S. (1997): Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces, Journal of Marketing (vol. 61), No. 3, pp. 38-53.

Andersen Legal (2001): Internet Privacy Survey - A Re-Survey of the Privacy Practices of Australia's Most Popular Websites, Sydney, 12 April 2001.

Andrews, Sarah (2002): Privacy and Human Rights 2002, London UK, Electronic Privacy Information Center, http://www.privacyinternational.org/survey/phr2002/

Ardissono, L. and Goy, A. (1999): Tailoring the Interaction with Users in Electronic Shops, User Modeling: Proceedings of the 7th International Conference, Banff, Canada. URL: http://www.cs.usask.ca/UM99/Proc/ardissono.pdf

Ardissono, Liliana and Goy, Anna (2000): Dynamic Generation of Adaptive Web Catalogs, Brusilivsky, Peter; Stock, Oliviero and Strappavara, Carlo, Adaptive Hypermedia and Adaptive Web-Based Systems (vol. 1892) pp. 5-16, Springer, Berlin.

Asonov, D. and Freytag, J. C. (2002): Almost Optimal Private Information Retrieval, Proceedings of the 2nd Workshop on Privacy Enhancing Technologies (PET '02), San Francisco.

AT&T (2002): AT&T Privacybird, http://www.privacybird.com/

Baldi, P.; Frasconi, P. and Smyth, P. (2003): Modeling the Internet and the Web. Probabilistic Methods and Algorithms., John Wiley & Sons, Chichester, UK.

BCG and Shop.Org (2002): The State of Retailing Online 5.0, June 2002, Press Release

BDSG (2003): Bundesdatenschutzgesetz, http://bundesrecht.juris.de/bundesrecht/bdsg_1990/htmltree.html

Beal, B. (2003): Analyzing the CRM analytics race, SearchCRM.com, http://searchcrm.techtarget.com/originalContent/0,289142,sid11_gci929770,00.html

Beck, L.L. (1980): A security mechanism for statistical databases, ACM Transactions on Database Systems (vol. 5), No. 5, pp. 316-328.

Behrens, L. (2001): Privacy and Security: The Hidden Growth Strategy, Gartner. 31 May 2001

Belanger, F.; Hiller, J.S. and Smith, W.J. (2002): Trustworthiness in electronic commerce: the role of privacy, security, and site attributes, Journal of Strategic Information Systems (vol. 11), pp. 245-270.

Bensberg, F. (2001): Web Log Mining als Instrument der Marketingforschung - Ein systemgestaltender Ansatz für internetbasierte Märkte [Web Log Mining as an Instrument for Marketing Research], Institut für Wirtschaftsinformatik, Westfälische Wilhelms-Universität Münster, Wiesbaden, Germany.

Berendt, B.; Günther, O. and Spiekermann, S. (2005): Privacy in E-Commerce: Stated Preferences vs. Actual Behavior, Communication of the ACM (vol. 48), No. 4, pp. 101-106.

Berendt, B.; Mobasher, B.; Spiliopoulou, M. and Wiltshire, J. (2001): Measuring the accuracy of sessionizers for web usage analysis, Proceedings of the Workshop on Web Mining at SIAM Data Mining Conference, Chicago, IL.

Berendt, B. and Spiliopoulou, M. (2000): Analysing navigation behaviour in Web sites integrating multiple information systems, VLDB Journal: Special Issue on Databases and the Web (vol. 9), No. 1, pp. 56-75.

Berendt, B. and Teltzrow, M. (2005): Addressing Users Privacy Concerns for Improving Personalization Quality: Towards an Integration of User Studies and Algorithm Evaluation, Mobasher, B. and Anand, S.S., Intelligent Techniques for Web Personalization, Berlin etc.: Springer. LNAI.

Berry, L.L. (1995): Relationship Marketing of Services - Growing Interest, Emerging Perspectives, Journal of the Academy of Marketing Science (vol. 23), No. 3, pp. 236-245.

Berthon, P.; Pitt, L.F. and Watson, R.T. (1996): The World Wide Web as an Advertising Medium, Journal of Advertising Research (vol. 36), No. 1, pp. 43-54.

Bhattacherjee, A. (2002): Individual trust in online firms: scale development and initial trust, Journal of Management Information Systems (vol. 19), No. 1, pp. 213-243.

Boyens, C. (2004): On privacy trade-offs in web-based services, Institute of Information Systems, Humboldt-Universität zu Berlin, Berlin.

Boyens, C.; Günther, O. and Teltzrow, M. (2002): Privacy Conflicts in CRM Services for Online Shops: A Case Study, Proceedings of the IEEE Workshop on Privacy, Security, and Data Mining, Maebashi, Japan.

Brodie, C.; Karat, C.-M. and Karat, J. (2004): How Personalization of an E-Commerce Website Affects Consumer Trust, Karat, J., Designing Personalized User Experience for eCommerce pp. 185-206, Kluwer Academic Publishers, Dordrecht, Netherlands.

Brusilovsky, P. (2001): Adaptive hypermedia, User Modeling and User-Adapted Interaction (vol. 11), No. 1-2, pp. 87-110.

Brusilovsky, P.; Kobsa, A. and Vassileva, J. (1998): Adaptive Hypertext and Hypermedia, Kluwer Academic Publishers, Dordrecht, Netherlands.

Büchner, A.; Mulvenna, M.; Anand, S. and Hughes, J. (1999): An Internet-enabled Knowledge Discovery Process, Proceedings of the 9th International Database Conference on Heterogeneous and Internet Databases, Hong Kong, China.

BWahlG (2005): Bundeswahlgesetz, http://bundesrecht.juris.de/bundesrecht/bwahlg/

Byrne, B. M. (1998): Structural equation modeling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications, and programming, Lawrence Erlbaum Associates, Mahwah, NJ.

CG&I-R (2001): Privacy Policies Critical to Online Consumer Trust, Columbus Group and Ipsos-Reid. Canadian Inter@ctive Reid Report

Chellappa, R.K. (2001): Consumers' Trust in Electronic Commerce Transactions: The Role of Perceived Privacy and Perceived Security, Working Paper, http://asura.usc.edu/~ram/rcf-papers/sec-priv.pdf

Chin, D. N. (1989): KNOME: Modeling What the User Knows in UC, Kobsa, A. and Wahlster, W., User Models in Dialog Systems pp. 74-107, Berlin, Heidelberg.

Chittaro, L. and Ranon, R. (2000): Adding Adaptive Features to Virtual Reality Interfaces for E-Commerce, Brusilivsky, P.; Stock, O. and Strappavara, C., Adaptive Hypermedia and Adaptive Web-Based Systems (vol. 1892) pp. 86-91, Springer, Berlin etc.

Cooley, R.; Mobasher, B. and Srivastava, J. (1999): Data preparation for mining World Wide Web browsing patterns, Journal of Knowledge and Information Systems (vol. 1), No. 1, pp. 5-32.

Cox, L.H. (1980): Suppression methodology and statistical disclosure control, Journal of the American Statistical Association (vol. 75), No. 370, pp. 377-385.

Cranor, L.; Langheinrich, M.; Marchiori, M.; Presler-Marshall, M. and Reagle, J. (2002): The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. W3C Recommendation 16 April 2002, http://www.w3.org/TR/P3P/

Cranor, L.; Reagle, J. and Ackerman, M. (1999): Beyond Concern: Understanding Net Users' Attitudes About Online Privacy, AT&T Labs - Research, Technical Report, http://www.research.att.com/resources/trs/TRs/99/99.4/99.4.3/report.htm

Crocker, D. and Overel, P. Augmented BNF for Syntax Specifications: ABNF, RFC2234, IETF, http://www.ietf.org/rfc/rfc2234.txt

Cronbach, L. J. (1951): Coefficient alpha and the internal structure of tests, Psychometrika (vol. 16), pp. 297-334.

Culnan, M.J. and Bies, R.J. (2003): Consumer Privacy: Balancing Economic and Justice Considerations, Journal of Social Issues (vol. 59), pp. 323-353.

Culnan, M.J. and Milne, G.R. (2001): The Culnan-Milne Survey on Consumers & Online Privacy Notices: Summary of Responses, Interagency Public Workshop: Get Noticed: Effective Financial Privacy Notices, Washington, D.C. URL: http://www.ftc.gov/bcp/workshops/glb/supporting/culnan-milne.pdf

Cutler, M. and Sterne, J. (2000): E-Metrics - Business Metrics for the New Economy, Netgenesis Corp., Technical Report, http://www.emetrics.org/articles/emetrics.pdf

Cyber Dialogue (2001): UCO Software To Address Retailers' $6.2 Billion Privacy Problem, http://www.cyberdialogue.com/news/releases/2001/11-07-uco-retail.pdf

Dai, H. and Mobasher, B. (2003): A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining, Proceedings of the International Conference on Internet Computing, Las Vegas, Nevada.

De Ruyter, K.; Wetzels, M. and Kleijnen, M. (2001): Customer adoption of e-services: an experimental study, International Journal of Service Industry Management (vol. 12), No. 2, pp. 184-207.

Debevc, M.; Meyer, B.; Donlagic, D. and Svecko, R. (1996): Design and evaluation of an adaptive icon toolbar, User Modeling and User-Adapted Interaction (vol. 6), No. 1, pp. 1-21.

Deloitte Touche Tohmatsu (2001): Dimension Data Privacy Survey, Canberra, http://www.didata.com.au/news/newsdetail.asp?ctID=174&sort=ctDateOfPubl%20DESC&filter=in&currentpage=1

Denning, D.E. (1982): Cryptography and Data Security, Addison-Wesley, Reading, MA.

Denning, D.E.; Denning, P.J. and Schwartz, M.D. (1979): The tracker: A threat to statistical database security, ACM Transactions on Database Systems (vol. 4), No. 1, pp. 76-79.

Department for Trade and Industry (2001): Informing Consumers about E-Commerce, London, Conducted by MORI, London: DTI, http://www.mori.com/polls/2001/pdf/dti-e-commerce.pdf

Deutsche Post Direkt GmbH (2004): One to one: mit guten Adressen die Richtigen erreichen, 5th of January, 2005, http://www.deutschepost.de//mlm.html/dpag/images/download/broschueren.Par.0107.File.pdf//download/broschueren/adressmanag_internet.pdf

Dobkin, D.; Jones, A.K. and Lipton, R.J. (1979): Secure databases: Protection against user influence, ACM Transactions on Database Systems (vol. 4), No. 1, pp. 97-106.

Domingo-Ferrer, J. and Herrera-Joancomarti, J. (1999): A privacy homomorphism allowing field operations on encrypted data, Jornades de Matematica Discreta i Algorismica, Universitat Politecnica de Catalunya.

Doney, P.M. and Cannon, J.P. (1997): An examination of the nature of trust in buyer-seller relationships, Journal of Marketing (vol. 61), No. 2, pp. 35-51.

Doyle, J.I. (2003): UCLA Internet Report, University of California, LA, February 2003, http://ccp.ucla.edu/pdf/UCLA-Internet-Report-Year-Three.pdf

Earp, J.B. and Baumer, D.C. (2003): Innovative Web Use to Learn about Consumer Behavior and Online Privacy, Communications of the ACM (vol. 46), No. 4, pp. 81-83.

Engel, J.F.; Kollat, D.T. and Blackwell, R.D. (1968): Consumer Behavior, Holt, Rinehart and Winston.

Ernst&Young (May 2004): P3P Dashboard Report, http://www.ey.com/global/download.nsf/US/P3P_Dashboard_-_May_2004/$file/E&YP3PDashboardMay2004.pdf

EU (1995): Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data, Official Journal of the European Communities, No. 23 November 1995 No L. 281, p. 31ff.

EU (2002): Directive 2002/58/EC of the European Parliament and of the Council

Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector.

Fellegi, I.P. (1972): On the question of statistical confidentiality, Journal of American Statistical Association (vol. 67), No. 337, pp. 7-18.

Fink, J.; Kobsa, A. and Nill, A. (1998): Adaptable and Adaptive Information Provision for All Users, Including Disabled and Elderly People, The New Review of Hypermedia and Multimedia (vol. 4), pp. 163-188.

Fischer-Hübner, S. (2001): IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms (vol. 1958 LNCS), Springer, Heidelberg-Berlin, Germany.

Fiutak, M. (2004): Cookies als Netzindikatoren ungeeignet, ZDNet. http://www.zdnet.de/news/tkomm/0,39023151,39126407,00.htm

Forrester (2001): Privacy issues inhibit online spending, Cambridge, MA, Forrester Research, Survey Summary, http://www.nua.ie/surveys?f=VS&art_id=905357259&rel=true

Foster, C. (2000): The Personalization Chain, Site Operations/Volume 3, Jupiter Communications

Fox, S.; Rainie, L.; Horrigan, J.; Lenhart, A.; Spooner, T. and Carter, C. (2000): Trust and Privacy Online: Why Americans Want to Rewrite the Rules, Washington, DC, The Pew Internet & American Life Project, http://www.pewinternet.org/pdfs/PIP_Trust_Privacy_Report.pdf

Friedman, B.; Howe, D.C. and Felten, E. (2002): Informed Consent in the Mozilla Browser: Implementing Value-Sensitive Design, Proceedings of the 35th Hawaii International Conference on System Sciences, Hawaii.

Fu, Y.; Sandhu, K. and Shih, M. (1999): Generalization-Based Approach to Clustering of Web Usage Sessions, Proceedings of the WebKDD Workshop, San Diego, CA.

Gallaugher, J.M. (2002): E-Commerce and the Undulating Distribution Channel,

Communications of the ACM (vol. 45), No. 7, pp. 89-95.

Gallo, R. and McAlister, J. (2003): The Top 50 Retailers, Retailforward, August 2003, Technical Report, http://www.retailforward.com

Gallup Organization (2001): Majority of E-mail Users Express Concern about Internet Privacy, Washington DC, June 28, 2001

Ganesan, S. (1994): Determinants of long-term orientation in buyer-seller relationships, Journal of Marketing (vol. 58), No. 2, pp. 1-19.

Garbarino, E. and Johnson, M.S. (1999): The Different Roles of Satisfaction, Trust, and Commitment in Customer Relationships, Journal of Marketing (vol. 63), No. 4, pp. 70-87.

GartnerG2 (2001): Privacy and Security: The Hidden Growth Strategy, Press Release, http://www4.gartner.com/5_about/press_releases/2001/pr20010807d.html

Gefen, D. (2000): E-commerce: the role of familiarity and trust, Omega: The International Journal of Management Science (vol. 28), No. 6, pp. 725-737.

Goersch, D. (2003): Multi-Channel Integration in the Retail of Physical Products, PhD School in Informatics, Copenhagen Business School, Copenhagen, Denmark.

Grabner-Kräuter, S. and Kaluscha, E.A. (2003): Empirical research in on-line trust: a review and critical assessment, International Journal of Human-Computer Studies (vol. 58), No. 6, pp. 783-812.

Gulati, R. and Garino, J. (2000): Get the Right Mix of Bricks and Clicks, Harvard Business Review (vol. 78), No. 3, pp. 107-114.

GVU (1998): GVU's 10th WWW User Survey, Graphics, Visualization and Usability Lab, Georgia Tech, 28th of May, 2000, http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/

Han, J. and Kamber, M. (2000): Data Mining: Concepts and Techniques, Gray, J., Ed,

Morgan Kaufmann Series in Data Management Systems.

Hansen, R. (2002): E-Commerce und Recht - Ein Leitfaden für Unternehmen, (Hrsg.), Hamann/Weidert, Ed, Erich-Schmidt-Verlag, Bielefeld.

Harper, J. and Singleton, S. (2001): With a Grain of Salt: What Consumer Privacy Surveys Don't Tell Us, Competitive Enterprises Institute, http://www.cei.org/PDFs/with_a_grain_of_salt.pdf

Harris Interactive (2000): A Survey of Consumer Privacy Attitudes and Behaviors, Rochester, NY, http://www.bbbonline.org/UnderstandingPrivacy/library/harrissummary.pdf

Harris Interactive (2001): Privacy Notices Research, Final Results, Initiative, Privacy Leadership. Rochester, NY

Harris Interactive (2003): Most People Are Privacy Pragmatists, Rochester NY.

Hawes, J.M.; Mast, K.W. and Swan, J.E. (1989): Trust earning perceptions of sellers and buyers, Journal of Personal Selling and Sales Management (vol. 9), No. 1, pp. 1-8.

Heer, J. and Chi, E.H (2002): Separating the Swarm: Categorization Methods for User Access Sessions on the Web, Proceedings of the ACM CHI 2002 Conference on Human Factors in Computing Systems, Minneapolis, MN.

Heijden, H. van der; Verhagen, T. and Creemers, M. (2001): Predicting Online Purchase Behavior: Replications and Test of Competing Models, Proceedings of the 34th Hawaii International Conference on System Sciences, Hawaii.

Hine, C. and Eve, J. (1998): Privacy in the Marketplace, The Information Society (vol. 14), No. 4, pp. 253-262. URL: http://taylorandfrancis.metapress.com/link.asp?id=033wvkeqd2weapjf.

Holynski, M. (1988): User-adaptive computer graphics, International Journal of Man-Machine Studies (vol. 29), No. 5, pp. 539-548.

Höök, K.; Karlgren, J.; Waern, A.; Dahlbäck, N.; Jansson, C.; Karlgren, K. and Lemaire, B. (1996): A glass box approach to adaptive hypermedia, User Modeling and User-Adapted Interaction (vol. 6), No. 2-3, pp. 157-184.

Howard, J.A. and Sheth, J.N. (1969): The Theory of Buying Behavior, John Wiley & Sons Inc., New York.

Hupprich, L. and Fan, J. (2004): Nielsen/Netratings Global Webwatch Finds Audiences of Top 10 Global Properties all Majority Male, December 2004, http://www.nielsen-netratings.com/pr/pr_020117_eratings.pdf

Interactive Policy (2002): Views on Data Protection, Questionnaire on the Implementation of the Data Protection Directive (95/46/EC). Results of Online Consultation, 20th of June - 15th of September 2002, Brussels

Ipsos Reid, Emailthatpays (2001): Canadians' Love Affair with Email Continues, http://www.ipsos-reid.com/media/content/displaypr.cfm?id_to_view=1345

Jameson, A.; Schäfer, R.; Simons, J. and Weis, Th. (1995): Adaptive Provision of Evaluation-Oriented Information: Tasks and Techniques, Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, San Mateo, CA.

Jarvenpaa, S. (1999): Consumer Trust in an Internet Store: A Cross-Cultural Validation, Journal of Computer-Mediated Communication (vol. 5), No. 2.

Jarvenpaa, S.; Tractinsky, N. and Vitale, M. (2000): Consumer trust in an Internet store, Information Technology and Management (vol. 1), No. 1-2, pp. 45-71.

Jennrich, R. I. and Sampson, P. F. (1966): Rotation for simple loadings, Psychometrika (vol. 31), pp. 313-323.

Joachims, Th.; Freitag, D. and Mitchell, T. (1997): WebWatcher: A Tour Guide for the World Wide Web, Proceedings of the International Joint Conference on Artificial Intelligence, Nagoya, Japan.

Joerding, T. (1999) (1999): A temporary user modeling approach for adaptive shopping on

the web, Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW, WWW-8, Toronto, Canada and UM99, Banff, Canada.

Jöreskog, K.G. and Sörbom, D. (1979): Advances in factor analysis and structural equation models, Cambridge, Mass.: Abt Books.

Jöreskog, K.G. and Sörbom, D. (1996): LISREL 8: Structural equation modeling with the SIMPLIS command language, Chicago, IL: Scientific Software International.

Jöreskog, K.G. and Sörbom, D. (2003): LISREL 8.54, SSI Central.

Jupiter Research Corporation (2001): CRM Moves into the Check-out Aisle - Multichannel Customer Integration Strategies, JUPT691271

Kaplan, C.; Fenwick, J. and Chen, J. (1993): Adaptive hypertext navigation based on user goals and context, User Modeling and User-Adapted Interaction (vol. 3), No. 3, pp. 193-220.

Karat, C.; Brodie, C.; Karat, J.; Vergo, J. and Alpert, S. (2003): Personalizing the User Experience on ibm.com, IBM Systems Journal (vol. 42), No. 2, pp. 686-701.

Karjoth, G.; Schunter, M. and Waidner, M. (2003): Platform for Enterprise Privacy Practices: Privacy-enabled Management of Customer Data, Proceedings of the 2nd Workshop on Privacy Enhancing Technologies, LNCS Volume 2482, San Francisco, USA.

KDNuggets (2005): Data Mining, Knowledge Discovery, Genomic Mining, Web Mining, http://www.kdnuggets.com

Kobsa, A. (2001): Generic User Modeling Systems, User Modeling and User-Adapted Interaction (vol. 11), No. 1-2, pp. 49-63.

Kobsa, A. (2002): Personalized Hypermedia and International Privacy, Communications of the ACM (vol. 45), No. 5, pp. 64-67. URL: http://www.ics.uci.edu/~kobsa/papers/2002-CACM-kobsa.pdf

Kobsa, A. (2003): A Component Architecture for Dynamically Managing Privacy Constraints in Personalized Web-Based Systems, Proceedings of the 3rd Workshop on Privacy Enhancing Technologies, Dresden, Germany.

Kobsa, A.; Koenemann, J. and Pohl, W. (2001): Personalized Hypermedia Presentation Techniques for Improving Customer Relationships, The Knowledge Engineering Review (vol. 16), No. 2, pp. 111-155.

Kobsa, A.; Müller, D. and Nill, A. (1994): KN-AHS: An Adaptive Hypertext Client of the User Modeling System BGP-MS, Proceedings of the 4th International Conference on User Modeling, Hyannis, MA.

Kobsa, A. and Schreck, J. (2003): Privacy through Pseudonymity in User-Adaptive Systems, ACM Transactions on Internet Technology (vol. 3), No. 2, pp. 149-183.

Kobsa, A. and Teltzrow, M. (2004): Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users' Data Sharing Behavior, Proceedings of the 4th Workshop on Privacy Enhancing Technologies, Toronto, Canada.

Kobsa, A. and Teltzrow, M. (2005): Impacts of Contextualized Communication of Privacy Practices and Personalization Benefits on Purchase Behavior and Perceived Quality of Recommendation, Proceedings of the Workshop "Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research" (IUI 2005), San Diego, CA.

Koch, R. (1998): The 80/20 Principle: The Secret of Achieving More With Less, Bantam Doubleday Dell Publishing, New York.

Kohavi, R. (2001): Mining E-Commerce Data: the Good, the Bad, and the Ugly, Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA.

Kohavi, R. (2003): Seminar on Computational Learning and Adaptation on Real-world Insights from Mining Retail E-Commerce Data, Stanford University, Center for the Study of Language and Information (CSLI), May 22, 2003,

http://ai.stanford.edu/users/ronnyk/realInsights.ppt

Kohavi, R. (2004): Front Line Internet Analytics at Amazon.com, Proceedings of the Emetrics Conference, June 2004, Santa Barbara, CA.

Kohavi, R. and Parekh, R. (2003): Ten Supplementary Analyses to Improve E-commerce Web Sites, Proceedings of the 5th ACM WebKDD 2003 Web Mining for E-Commerce Workshop "Webmining as a Premise to Effective and Intelligent Web Applications", Washington, D.C.

Konstan, J. A.; Miller, B. N.; Maltz, D.; Herlocker, J. L.; Gordon, L. R. and Riedl, J. (1997): GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM (vol. 40), No. 3, pp. 77-87.

Kosala, R. and Blockeel, H. (2000): Web mining research: A survey., SIGKDD Explorations (vol. 2), No. 1, pp. 1-15.

Koufaris, M. and Hampton-Sosa, W. (2002): Customer trust online: examining the role of the experience with the Web-site, CIS Working Paper Series, Zicklin School of Business, Baruch College, New York, NY, http://cisnet.baruch.cuny.edu/papers/cis200205.pdf

Kozierok, R. and Maes, P. (1993): A Learning Interface Agent for Scheduling Meetings, Proceedings of the International Workshop on Intelligent User Interfaces, Orlando, FL.

Krogsaeter, M.; Oppermann, R. and Thomas, C. G. (1994): A User Interface Integrating Adaptability and Adaptivity, Oppermann, R., Adaptive User Support pp. 97-125, Lawrence Erlbaum Associates, Inc.

Küpper, D. and Kobsa, A. (1999): User-Tailored Plan Generation, Proceedings of the User Modeling 7th International Conference, Banff, Canada. URL: http://www.ics.uci.edu/~kobsa/papers/1999-UM99-kobsa.pdf

Lamm, S.E. ; Reed, D.A. and Scullin, W.H. (1996): Real-Time Geographic Visualization of World Wide Web Traffic, Computer Networks (vol. 28), No. 7-11, pp. 1457-1468.

Lederer, S.; Dey, A. and Mankoff, J. (2002): A Conceptual Model and Metaphor of Everyday Privacy in Ubiquitous Computing, Research, Intel. Technical Report IRB-TR-02-017, http://www.intel-research.net/Publications/Berkeley/120520020944_107.pdf

Lee, J.; Podlaseck, M.; Schonberg, E. and Hoch, R. (2001): Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandizing, Data Mining and Knowledge Discovery (vol. 5), No. 1/2, pp. 59-84.

LeFevre, K.; Agrawal, R.; Ercegovac, V.; Ramakrishnan, R.; Xu, Y. and DeWitt, D.J. (2004): Limiting Disclosure in Hippocratic Databases, Proceedings of the 30th International Conference on Very Large Data Bases, Toronto, Canada.

Lefons, E.; Silvestri, A. and Tangorra, F. (1983): An analytical approach to statistical databases, Proceedings of the 9th International Conference on Very Large Databases.

Lesh, N.; Rich, C. and Sidner, C.L. (1999): Using plan recognition in human-computer collaboration, Proceedings of the User Modeling 7th International Conference, Banff, Canada.

Lieberman, H. (1995): Letizia: An Agent that Assists Web Browsing, Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.

Likert, R. (1932): A Technique for the Measurement of Attitudes, Summers, G., Attitude Measurement pp. 149-158, Rand McNally&Company, Chicago, IL.

Lorenz, M. O. (1905): Methods for Measuring the Concentration of Wealth, American Statistical Association (vol. 9), pp. 209-219.

Lynch, P.J. and Horton, S. (2001): Web Style Guide, Yale University. URL: http://www.webstyleguide.com/site/index.html

Mabley, Kevin (2000): Privacy vs. Personalization: Part III, Cyber Dialogue, Inc., http://www.cyberdialogue.com/library/pdfs/wp-cd-2000-privacy.pdf

Maes, P. (1994): Agents that Reduce Work and Information Overload, Communications of the ACM (vol. 37), No. 7, pp. 31-40.

Malacinski, A.; Dominick, S. and Hatrick, T. (2001): Measuring Web Traffic, IBM. http://www-106.ibm.com/developerworks/web/library/wa-mwt1

Markillie, P. (2004): A Perfect Market: A Survey of E-Commerce, Economist (vol. 371), No. 8375, pp. 3-5.

Melissa Data (2004): IP2Location - Know where your web visitors are located, 5th of January, 2005, http://www.melissadata.com/lookups/ip2location.asp

Miglautsch, J.R. (2000): Thoughts on RFM scoring, Journal of Database Marketing (vol. 8), No. 1, pp. 67-72.

Mitchell, T.; Caruana, R.; Freitag, D.; McDermott, J. and Zabowski, D. (1994): Experience with a Learning Personal Assistant, Communications of the ACM (vol. 37), No. 7, pp. 81-91.

Mobasher, B.; Cooley, R. and Srivastava, J. (2000a): Automatic personalization based on web usage mining, Communications of the ACM (vol. 43), No. 8, pp. 142-151.

Mobasher, B.; Dai, H.; Luo, T. and Nakagawa, M. (2002): Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization., Data Mining and Knowledge Discovery (vol. 6), No. 1, pp. 61-82.

Mobasher, B.; Dai, H.; Luo, T.; Sung, Y. and Zhu, J. (2000b): Integrating Web Usage and Content Mining for More Effective Personalization, Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000), Greenwich, UK.

Moe, W. (2001): Buying, Searching, or Browsing: Differentiating between Online Shoppers In-Store Navigational Clickstream, Journal of Consumer Psychology (vol. 13), No. 1&2.

Moe, W. and Fader, P. (2000): Capturing Evolving Visit Behavior in Clickstream Data.,

The Wharton School, Working Paper, http://www-marketing.wharton.upenn.edu/ideas/pdf/00-003.pdf

Monticino, M. (1998): Web-Analysis: stripping away the hype, IEEE Computer (vol. 31), No. 12, pp. 130-132.

Morgan, R.M. and Hunt, S.D. (1994): The Commitment-Trust Theory of Relationship Marketing, Journal of Marketing (vol. 58), No. 3, pp. 20-38.

Mulvenna, M. D.; Anand, S.S. and Buchner, A.G. (2000): Personalization on the net using web mining, Communications of the ACM (vol. 43), No. 8, pp. 123-125.

Neiling, M. (2004): Identifizierung von Realwelt-Objekten in multiplen Datenbanken, Brandenburgische Technische Universität Cottbus. URL: http://www.ub.tu-cottbus.de/hss/diss/fak1/neiling_m/pdf/diss_neiling.pdf

Nelson, P.J. (1974): Advertising as Information, Journal of Political Economy (vol. 82), No. 4, pp. 729-754.

Newcombe, H.B.; Fair, M.E. and Lalonde, P. (1992): The use of names for linking personal records, Journal of the American Statistical Association (vol. 87), pp. 1193-1204.

Nicosia, F.M. (1966): Consumer Decision Processes, Prentice-Hall, Englewood Cliffs, NJ.

Olsen, S. (2000): Geographic tracking raises opportunities, fears, CNET News.com. 28th of April, 2005, http://news.com.com/2100-1023-248274.html?legacy=cnet

Omwando, H. (2002): Choosing the Right Retail Channel Strategy, Forrester Research, Tech Strategy Report

Otto, J.R. and Chung, Q.B. (2000): A Framework for Cyber Enhanced Retailing, Electronic Markets (vol. 10), No. 3, pp. 185-191.

P3P (2002): W3C Platform for Privacy Preferences Initiative, Version 1.0 as of April 2002, http://www.w3.org/P3P

Palen, L. and Dourish, P. (2002): Unpacking “Privacy” for a Networked World, Proceedings of the Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA.

Patrick, A.S. and Kenny, S. (2003): From Privacy Legislation to Interface Design: Implementing Information Privacy in Human-Computer Interfaces, Dingledine, R., Privacy Enhancing Technologies (vol. LNCS 2760) pp. 107-124, Springer Verlag, Heidelberg, Germany. URL: http://www.andrewpatrick.ca/legint/pet-workshop-patrick-kenny.pdf

Pavlou, P.A. (2003): Consumer acceptance of electronic commerce - integrating trust and risk with the technology acceptance model, International Journal of Electronic Commerce (vol. 7), No. 3, pp. 69-103.

Pazzani, M. and Billsus, D. (1997): Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning (vol. 27), pp. 313-331.

Pepper, P. (2003): Funktionale Programmierung in OPAL, ML, HASKELL und Gofer., 2nd edition. ed., Springer.

Perkowitz, M. and Etzioni, O. (1998): Adaptive web pages: Automatically synthesizing web pages., Proceedings of the Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin.

Perkowitz, M. and Etzioni, O. (2000): Adaptive web sites, Communications of the ACM (vol. 43), No. 8, pp. 152-158.

Personalization Consortium (2000): Personalization & Privacy Survey, Edgewater Place, MA, Personalization Consortium, http://www.personalization.org/SurveyResults.pdf

Phau, I. and Poon, S.M. (2000): Factors Influencing the Types of Products and Services Purchased over the Internet, Internet Research (vol. 10), No. 2, pp. 102-113.

Pohle, C. and Spiliopoulou, M. (2002): Building and Exploiting Ad Hoc Concept Hierarchies for Web Log Analysis, Proceedings of the 4th International Conference Data Warehousing and Knowledge Discovery (DaWaK), Aix-en-Provence, France.

Popp, H. and Lödel, D. (1996): Fuzzy techniques and user modeling in sales assistants, User Modeling and User-Adapted Interaction (vol. 5), No. 3-4, pp. 349-370.

Privacy & American Business (1999): Personalized Marketing and Privacy on The Net: What Consumers Want, November 1999, http://www.pandab.org/doubleclicksummary.html

Reinartz, W.J. and Kumar, V. (2003): The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration (vol. 67), No. 1, pp. 77-99.

Resnick, P. and Varian, H. R. (1997): Recommender Systems, Communications of the ACM (vol. 40), No. 3, pp. 56-58.

Responsys.com (2000): Online Marketers Have Little Confidence in Self-Regulation of Internet Privacy. Sponsored by Millard Brown IntelliQuest, Palo Alto, CA.

Rivest, R.; Adleman, L. and Dertouzos, M. (1978): On Data Banks and Privacy Homomorphisms, DeMillo, R.A., Foundations of Secure Computation, Academic Press, New York.

Rosen, K.T. and Howard, A.L. (2000): E-Retail: Gold Rush or Fool's Gold?, California Management Review (vol. 42), No. 3, pp. 72-100.

Rotenberg, Marc (2001): The Privacy Law Sourcebook 2001: United States Law, International Law, and Recent Developments, EPIC, Washington, DC.

Roy Morgan Research (2001): Privacy and the Community, Sydney, Prepared for the Office of the Federal Privacy Commissioner, http://www.privacy.gov.au/publications/rcommunity.html

Sakagami, H.; Kamba, T.; Sugiura, A. and Koseki, Y. (1998): Effective personalization of push-type systems: visualizing information freshness, Proceedings of the 7th World Wide Web Conference, Brisbane, Australia.

SAP AG (2001): CRM Analytics, Presentation Slides, part of my SAP CRM

Sarwar, B.; Karypis, G.; Konstan, J. and Riedl, J. (2000): Analysis of Recommendation Algorithms for E-Commerce., Proceedings of the 1st ACM Conference on Electronic Commerce, San Diego, CA.

Schneemann, K. (2003): Who clicks Who? Online Reichweiten Monitor 2003 II, G+J Electronic Media Sales GmbH. Die Blaue Reihe, http://www.gujmedia.de/_components/sidebars/mediaservice_sidebar/EMS_ORM_2003_II.pdf

Schwickert, A. C. (2001): Controlling-Kennzahlen für Web Sites, 5. Internationale Tagung Wirtschaftsinformatik "Information Age Economy", Augsburg, Germany.

Shahabi, C.; Zarkesh, A.M.; Adibi, J. and Shah, V. (1997): Knowledge Discovery from User's Web-page Navigation, Proceedings of the 7th IEEE International Conference On Research Issues in Data Engineering.

Shardanand, U. and Maes, P. (1995): Social Information Filtering: Algorithms for Automating 'Word of Mouth', Proceedings of CHI-95 (Human Factors in Computing).

Shern, S. (2001): Global Online Retailing, Ernst&Young. http://www.ey.com/global/download.nsf/International/Global_Online_Retailing_-_Special_Report_2001/$file/GOR_2001.pdf

Shneiderman, B. (2000): Universal Usability, Communications of the ACM (vol. 43), No. 5, pp. 85-91.

Shneiderman, B. and Plaisant, C. (2004): Designing the User Interface, 4th. ed., Pearson Addison Wesley.

Shop.org and Forrester Research (2004): The State of Retailing Online 7.0, http://www.shop.org/research/SRO7/SRO7main.asp

Silverstein, M.J.; Sirkin, H.L. and Stanger, P. (2002): The State of Retailing Online 5.0, BCG and Shop.org. 28th of April, 2005, http://www.bcg.com/publications/publications_search_results.jsp?PUBID=755

Sitkin, S.B. and Weingart, L.R. (1995): Determinants of risky decision-making behavior: a test of the mediating role of risk perceptions and propensity, Academy of Management Journal (vol. 38), No. 6, pp. 1573-1592.

Sommerville, I. (2004): Software Engineering, 8th. ed., Pearson.

Specht, M. (1998): Empirical Evaluation of Adaptive Annotation in Hypermedia, Proceedings of the ED-MEDIA98, Freiburg, Germany.

Spiekermann, S.; Grossklags, J. and Berendt, B. (2001): E-privacy in 2nd Generation E-Commerce: Privacy Preferences versus Actual Behavior, Proceedings of the 3rd ACM Conference on Electronic Commerce, Tampa, FL.

Spiliopoulou, M. (1999): The laborious way from data mining to Web mining, International Journal of Computer Systems, Science & Engineering, Special Issue on "Semantics of the Web" (vol. 14), pp. 113-126.

Spiliopoulou, M. (2000): Web usage mining for Web site evaluation, Communications of the ACM (vol. 43), No. 8, pp. 127-134.

Spiliopoulou, M. and Berendt, B. (2001): Kontrolle der Präsentation und Vermarktung von Gütern im WWW anhand von Data-Mining-Techniken, Hippner, H.; Küsters, M.; Meyer, M. and Wilde, K.D., Handbuch Data Mining - Knowledge Discovery in Databases pp. 855-873, Vieweg, Wiesbaden, Germany.

Spiliopoulou, M. and Faulstich, L. C. (1999): WUM: A Tool for Web Utilization Analysis, Extended Version of Proceedings Workshop WebDB'98 at the International Conference on Extending DataBase Technology (LNCS 1590, pp. 184-203).

Spiliopoulou, M. and Faulstich, L.C. (1998): WUM: A Web utilization miner, Proceedings of the Workshop WebDB98 at the International Conference on Extending Database Technology, Valencia, Spain.

Spiliopoulou, M.; Mobasher, B. and Berendt, B. (2002a): Web Usage Mining for E-Business Applications, Tutorial at the 13th European Conference on Machine Learning (ECML'02) / 6th European Conference on Principles and Practice of

Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland.

Spiliopoulou, M.; Mobasher, B.; Berendt, B. and Nakagawa, M. (2003): A framework for the evaluation of session reconstruction heuristics in Web-usage analysis, INFORMS Journal on Computing (vol. 15), pp. 171-190.

Spiliopoulou, M. and Pohle, C. (2001): Data mining for measuring and improving the success of web sites, Data Mining and Knowledge Discovery (vol. 5), No. 1-2, pp. 85-114.

Spiliopoulou, M.; Pohle, C. and Teltzrow, M. (2002b): Modelling Web Site Usage with Sequences of Goal-Oriented Tasks, Multikonferenz Wirtschaftsinformatik, Nürnberg, Germany.

Srivastava, J.; Cooley, R.; Deshpande, M. and Tan, P.-N. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations (vol. 1), No. 2, pp. 12-23.

Srivastava, J.; Prasanna, D. and Kumar, V. (2002): Web Mining: Accomplishments & Future Directions, National Science Foundation Workshop on Next Generation Data Mining (NGDM'02), Baltimore, Maryland.

Stallings, W. (1999): Cryptography and Network Security: Principles and Practice, Prentice Hall.

Steinfield, C. (2002): Understanding Click and Mortar E-Commerce Approaches: A Conceptual Framework and Research Agenda, Journal of Interactive Advertising (vol. 2), No. 2.

Stone, M.; Hobbs, M. and Khaleeli, M. (2002): Multichannel customer management: the benefits and challenges, Journal of Database Marketing (vol. 10), No. 1, pp. 39-53.

Subramaniam, C.; Shaw, M.J. and Gardner, D.M. (2000): Product Marketing and Channel Management in Electronic Commerce, Information Systems Frontier (vol. 1), No. 4, pp. 363-378.

Sweeney, L. (2001): Computational Disclosure Control: A Primer on Data Privacy Protection, Massachusetts Institute of Technology, Cambridge. URL: http://www.swiss.ai.mit.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf

Sweeney, L. (2002): K-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems (vol. 10), No. 7, pp. 557-570.

Swerdlow, F.S.; Deeks, J. and Cassar, K. (2002): In-Store Pickup Implement Last Mile Policies to Optimize Technology Investments., Jupiter Research, Tech Report, vol. 10

TDDSG (2001): Gesetz über den Datenschutz bei Telediensten (Teledienstedatenschutzgesetz - TDDSG) [Teleservices Data Protection Law]. Artikel 3 des Gesetzes über rechtliche Rahmenbedingungen für den elektronischen Geschäftsverkehr (Elektronischer Geschäftsverkehr-Gesetz – EGG) vom 14 Dezember 2001, BGBl I, 3721.

Tedeschi, B. (2001): Web Retailers Add In-Store Pick-Up, 360Commerce, 5th of January, 2005, http://www.360commerce.com/display.php?tid=129

Teltzrow, M. and Berendt, B. (2003): Web-Usage-Based Success Metrics for Multi-Channel Businesses, Proceedings of the 5th ACM WebKDD 2003 Web Mining for E-Commerce Workshop, Washington, D.C.

Teltzrow, M.; Berendt, B. and Günther, O. (2003a): Consumer Behaviour at Multi-Channel Retailers, Proceedings of the 4th IBM eBusiness Conference, University of Surrey, School of Management, UK. URL: http://www.wiwi.hu-berlin.de/~teltzrow/MCBehavior.pdf

Teltzrow, M.; Berendt, B. and Günther, O. (2004a): Ein Kennzahlensystem für Mehrkanalhändler, Multikonferenz Wirtschaftsinformatik (MKWI), Essen, Germany.

Teltzrow, M. and Günther, O. (2001): eCRM: Konzeption und Möglichkeiten zur Effizienzmessung [eCRM: Concepts and Options for Performance Evaluation], Praxis der Wirtschaftsinformatik (vol. 221), pp. 16-26.

Teltzrow, M. and Günther, O. (2003): Web Usage Metrics for Multi-Channel Retailers, Proceedings of the 4th International Conference EC-Web, Prague, Czech Republic.

Teltzrow, M.; Günther, O. and Pohle, C. (2003b): Analyzing Consumer Behavior at Retailers with Hybrid Distribution Channels - A Trust Perspective, ACM International Conference Proceedings, Proc. 5th International Conference of Electronic Commerce, Pittsburgh, Pennsylvania.

Teltzrow, M. and Kobsa, A. (2003): Impacts of User Privacy Preferences on Personalized Systems - a Comparative Study, CHI-2003 Workshop "Designing Personalized User Experiences for eCommerce: Theory, Methods, and Research", Fort Lauderdale, FL.

Teltzrow, M. and Kobsa, A. (2004a): Communication of Privacy and Personalization in E-Business, Proceedings of the Workshop “WHOLES: A Multiple View of Individual Privacy in a Networked World”, Stockholm, Sweden. URL: http://www.ics.uci.edu/~kobsa/papers/2004-WHOLES-kobsa.pdf

Teltzrow, M. and Kobsa, A. (2004b): Impacts of User Privacy Preferences on Personalized Systems: a Comparative Study, Karat, C.-M.; Blom, J. and Karat, J., Designing Personalized User Experiences for eCommerce pp. 315-332, Kluwer Academic Publishers, Dordrecht, Netherlands.

Teltzrow, M.; Preibusch, S. and Berendt, B. (2004b): SIMT - A Privacy-Preserving Web Metrics Tool, Proceedings of the IEEE Conference on Electronic Commerce (CEC04), San Diego.

Teo, Hock-Hai ; Wan, W. and Li, L. (2004): Volunteering Personal Information on the Internet: Effects of Reputation, Privacy Initiatives, and Reward on Online Consumer Behavior, Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii, USA.

Torra, V. (2000): Re-identifying individuals using OWA operators, Proceedings of the 6th International Conference on Soft Computing, Iizuka, Fukuoka, Japan.

UMR (2001): Privacy Concerns Loom Large. Conducted for the Privacy Commissioner of New Zealand, Survey summary, Auckland: PC of New Zealand. URL: http://www.privacy.org.nz/privword/42pr.html

USA Today (2003): Internet sales soar over Thanksgiving weekend, Money, New York, 2nd of December, 2003, http://www.usatoday.com/money/industries/retail/2003-12-02-internet-sales_x.htm

van Duyne, D. K.; Landay, J. A. and Hong, J. I. (2002): The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience, Addison-Wesley, Boston.

W3C (1995): Logging Control In W3C httpd, http://www.w3.org/Daemon/User/Config/Logging.html

Weichert, T. (2004): Geomarketing und Datenschutz, Symposium "Living by numbers", Düsseldorf, Germany. URL: http://www.datenschutzzentrum.de/wirtschaft/vortrag_geomarketing.htm

Weigend, A. (2003): Analyzing Customer Behavior at Amazon.com (Invited Talk), Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, Washington, DC.

Wilde, E. (2003): Wilde's WWW - Technical Foundations of the World Wide Web, 2nd. ed., Springer, New York.

Willenborg, L. and Waal, T. de (2001): Elements of Statistical Disclosure Control, Addison Wesley.

Winkler, W.E. (1995): Matching and record linkage, Cox, B.G., Business Survey Methods pp. 355-384, J. Wiley, New York.

Wright, A.A. and Lynch, J.G.Jr. (1995): Communications effects of advertising versus direct experience when both search and experience attributes are present, Journal of Consumer Research (vol. 21), No. 3, pp. 708-718.

Yu, C.T. and Chin, F.Y. (1977): A study on the protection of statistical databases, Proceedings of the ACM SIGMOD International Conference of the Management of Data, Toronto, Canada.

Zaiane, O.R.; Xin, M. and Han, J. (1998): Discovering web access patterns and trends by applying OLAP and data mining technology on web logs, Proceedings of Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA.

Appendix

Appendix to Chapter 2

Data tables

Scale and Items	Factor Loadings	Source
Willingness to Buy		
WTB1. How likely is it that you would consider purchasing from this store in the short term?	0.88	[Heijden, et al., 2001], (based on [Jarvenpaa, 1999; Jarvenpaa, et al., 2000])
WTB2. How likely is it that you would consider purchasing from this store in the long term?	0.85	
WTB33. For this purchase, how likely is it that you buy from this store?	0.69	
Perceived Physical Store Size		
PS1. This retailer's stores are spread all over the country.	0.91	Modified items according to [Doney and Cannon, 1997; Jarvenpaa, et al., 2000]
PS2. This retailer's store network is relatively small in its home market. [reverse]	0.79	
PS3. The retailers' stores belong to a large company.	0.61	
Perceived Physical Store Reputation		
PR1. This retailer's stores are well known.	0.80	[Doney and Cannon, 1997]
PR2. This retailer's stores have a bad reputation in the market. [reverse]	0.94	
PR3. This retailer's stores have a good reputation.	0.88	
Store Trustworthiness		[Doney and Cannon, 1997; Heijden, et al., 2001; Jarvenpaa,
TR1. This e-shop is trustworthy.	-0.52	
TR2. This e-shop keeps its commitments and promises.	-0.58	

TR3. The experiences with this e-shop met my expectations.	-0.76	1999; Jarvenpaa, et al., 2000; Koufaris and Hampton-Sosa, 2002; Pavlou, 2003]
Privacy		
PRI1. I have no concerns transmitting personal data to this e-shop.	0.90	
PRI2. This e-shop handles my personal data responsibly.	0.93	[Chellappa, 2001]
PRI3. My personal data are in good hands at this retailer.	0.89	
Risk perception		
RP1. What is the likelihood of your making a good bargain by buying from this store through the Internet? (very unlikely – very likely)	0.69	[Jarvenpaa, 1999; Jarvenpaa, et al., 2000; Sitkin and Weingart, 1995]
RP2. How would you characterize the decision to buy a product through this Web site? (high potential for loss – high potential for gain) [reverse]	0.84	
RP3. How would you characterize the risk to purchase at this e-shop? (very low risk, very high risk) [reverse]	0.71	

Table 0-1: Scales, items and sources

Item	Component					
	1	2	3	4	5	6
WTB1		0.88				
WTB2		0.85				
WTB3		0.69				
PS1				0.91		
PS2				0.79		
PS3				0.61		
PR1	0.80					
PR2	0.94					
PR3	0.88					
TR1						-0.52
TR2						-0.58
TR3						-0.76
PRI1			0.90			
PRI2			0.93			
PRI3			0.89			
RP1					0.69	
RP2					0.84	
RP3					0.71	

Table 0-2: Pattern matrix of the rotated six factor solution

Note. Extraction Method: Principal Component Analysis, Rotation Method: Oblimin with Kaiser Normalization. Loadings below .3 are omitted; loadings above .55 are in bold face.

Lisrel output

Model 1 (n=524)

Observed Variables

Willingness to Buy	Perceived Size	Perceived Reputation	Trust	Privacy	Risk Perception
WTB1	PS1	PR1	TR1	PRI1	RP1
WTB2	PS2	PR2	TR2	PRI2	RP2
WTB3	PS3	PR3	TR3	PRI3	RP3

Sample Size: 524

Latent Variables: TR PS PR PRI

Relationships

TR1 = TR

TR2 = TR

TR3 = TR

PS1 = PS

PS2 = PS

PS3 = PS

PR1 = PR

PR2 = PR

PR3 = PR

PRI1 = PRI

PRI2 = PRI

PRI3 = PRI

TR = PS PR PRI

Correlation Matrix

	TR1	TR2	TR3	PS1	PS2	PS3
TR1	1.000					
TR2	0.712	1.000				
TR3	0.648	0.730	1.000			
PS1	0.354	0.308	0.300	1.000		
PS2	0.437	0.404	0.392	0.618	1.000	
PS3	0.391	0.402	0.347	0.430	0.519	1.000
PR1	0.526	0.560	0.520	0.391	0.461	0.433
PR2	0.565	0.526	0.463	0.330	0.490	0.389
PR3	0.567	0.612	0.499	0.301	0.413	0.427
PRI1	0.518	0.632	0.447	0.183	0.283	0.301
PRI2	0.511	0.661	0.512	0.240	0.327	0.292
PRI3	0.537	0.632	0.515	0.264	0.316	0.292

Correlation Matrix

	PR1	PR2	PR3	PRI1	PRI2	PRI3
PR1	1.000					
PR2	0.756	1.000				
PR3	0.743	0.803	1.000			
PRI1	0.335	0.375	0.404	1.000		
PRI2	0.337	0.351	0.368	0.919	1.000	
PRI3	0.358	0.339	0.407	0.873	0.927	1.000

Parameter Specifications

LAMBDA-Y	
TR	

TR1	0
TR2	1
TR3	2

LAMBDA-X			
	PS	PR	PRI
PS1	3	0	0
PS2	4	0	0
PS3	5	0	0
PR1	0	6	0
PR2	0	7	0
PR3	0	8	0
PRI1	0	0	9
PRI2	0	0	10
PRI3	0	0	11

GAMMA			
	PS	PR	PRI
TR	12	13	14
PHI			
	PS	PR	PRI
PS	0		
PR	15	0	
PRI	16	17	0
PSI			
	TR		
	18		

THETA-EPS			
	TR1	TR2	TR3
	19	20	21

THETA-DELTA									
	PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
	22	23	24	25	26	27	28	29	30

Number of Iterations = 11

LISREL Estimates (Weighted Least Squares)

LAMBDA-Y	
	TR
TR1	0.888
TR2	0.953
	(0.028)
	34.500
TR3	0.843
	(0.031)
	27.167

LAMBDA-X			
	PS	PR	PRI
PS1	0.693	-	-
	(0.041)		
	17.095		
PS2	0.868		

	(0.028)		
	31.240		
PS3	0.750	-	-
	(0.029)		
	26.302		
PR1	-	0.941	-
		(0.017)	
		55.718	
PR2	-	0.960	-
		(0.015)	
		62.532	
PR3	-	0.959	
		(0.014)	
		66.471	
PRI1	-	-	0.968
			(0.012)
			83.703
PRI2	-	-	0.991
			(0.008)
			131.542
PRI3	-	-	1.000
			0.010
			96.539

GAMMA			
	PS	PR	PRI
TR	0.166	0.413	0.461

	(0.060)	(0.060)	(0.046)
	2.772	6.897	10.039

Covariance Matrix of ETA and KSI

	TR	PS	PR	PRI
TR	1.000			
PS	0.703	1.000		
PR	0.774	0.708	1.000	
PRI	0.767	0.553	0.529	1.000

PHI			
	PS	PR	PRI
PS	1.000		
PR	0.708	1.000	
	(0.036)		
	19.423		
PRI	0.533	0.529	1.000
	(0.042)	(0.040)	
	12.554	13.074	
PSI			
	TR		
	0.210		
	(0.033)		
	6.423		

Squared Multiple Correlations for Structural Equations

TR
0.790

Squared Multiple Correlations for Reduced Form

TR
0.790

THETA-EPS		
TR1	TR2	TR3
0.211	0.092	0.289
(0.060)	(0.051)	(0.061)
3.532	1.790	4.758

Squared Multiple Correlations for Y - Variables

TR1	TR2	TR3
0.789	0.908	0.711

THETA-DELTA								
PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.520	0.247	0.437	0.114	0.078	0.080	0.063	0.018	0.001
(0.071)	(0.065)	(0.061)	(0.054)	(0.053)	(0.052)	(0.049)	(0.046)	(0.048)
7.298	3.797	7.138	2.107	1.477	1.544	1.275	0.386	0.011

Squared Multiple Correlations for X - Variables

PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.480	0.753	0.563	0.886	0.922	0.920	0.937	0.982	0.999

Goodness of Fit Statistics

Degrees of Freedom = 48

Minimum Fit Function Chi-Square = 96.167 (P = 0.000)

Estimated Non-centrality Parameter (NCP) = 48.167

90 Percent Confidence Interval for NCP = (24.032 ; 80.082)

Minimum Fit Function Value = 0.184

Population Discrepancy Function Value (F0) = 0.0921

90 Percent Confidence Interval for F0 = (0.0459 ; 0.153)

Root Mean Square Error of Approximation (RMSEA) = 0.0438

90 Percent Confidence Interval for RMSEA = (0.0309 ; 0.0565)

P-Value for Test of Close Fit (RMSEA < 0.05) = 0.778

Expected Cross-Validation Index (ECVI) = 0.299

90 Percent Confidence Interval for ECVI = (0.252 ; 0.360)

ECVI for Saturated Model = 0.298

ECVI for Independence Model = 21.021

Chi-Square for Independence Model with 66 Degrees of Freedom = 10969.722

Independence AIC = 10993.722

Model AIC = 156.167

Saturated AIC = 156.000

Independence CAIC = 11056.860

Model CAIC = 314.012

Saturated CAIC = 566.396

Normed Fit Index (NFI) = 0.991

Non-Normed Fit Index (NNFI) = 0.994

Parsimony Normed Fit Index (PNFI) = 0.721

Comparative Fit Index (CFI) = 0.996

Incremental Fit Index (IFI) = 0.996

Relative Fit Index (RFI) = 0.988

Critical N (CN) = 401.722

Root Mean Square Residual (RMR) = 0.114

Standardized RMR = 0.114

Goodness of Fit Index (GFI) = 0.994

Adjusted Goodness of Fit Index (AGFI) = 0.991

Parsimony Goodness of Fit Index (PGFI) = 0.612

Model 2 (n=524)

Observed Variables

Willingness to Buy	Perceived Size	Perceived Reputation	Trust	Privacy	Risk Perception
WTB1	PS1	PR1	TR1	PRI1	RP1
WTB2	PS2	PR2	TR2	PRI2	RP2
WTB3	PS3	PR3	TR3	PRI3	RP3

Sample Size: 524

Latent Variables: TR PS PR PRI

Relationships

TR1 = TR

TR2 = TR

TR3 = TR

PS1 = PS

PS2 = PS

PS3 = PS

PR1 = PR

PR2 = PR

PR3 = PR

PRI1 = PRI

PRI2 = PRI

PRI3 = PRI

TR = PS PR PRI

Correlation Matrix

	TR1	TR2	TR3	PS1	PS2	PS3
TR1	1.000					
TR2	0.685	1.000				
TR3	0.624	0.666	1.000			
PS1	0.372	0.332	0.362	1.000		
PS2	0.454	0.351	0.325	0.583	1.000	
PS3	0.483	0.433	0.363	0.477	0.650	1.000
PR1	0.564	0.472	0.544	0.450	0.456	0.440
PR2	0.550	0.451	0.479	0.359	0.395	0.491
PR3	0.560	0.544	0.498	0.396	0.446	0.496
PRI1	0.527	0.669	0.463	0.318	0.295	0.465
PRI2	0.540	0.679	0.473	0.344	0.349	0.489
PRI3	0.555	0.685	0.467	0.347	0.319	0.507

Correlation Matrix

	PR1	PR2	PR3	PRI1	PRI2	PRI3
PR1	1.000					
PR2	0.713	1.000				
PR3	0.697	0.765	1.000			
PRI1	0.384	0.362	0.432	1.000		
PRI2	0.394	0.354	0.449	0.912	1.000	
PRI3	0.398	0.397	0.434	0.903	0.919	1.000

Parameter Specifications

LAMBDA-Y	
TR	
TR1	0
TR2	1
TR3	2

LAMBDA-X			
	PS	PR	PRI
PS1	3	0	0
PS2	4	0	0
PS3	5	0	0
PR1	0	6	0
PR2	0	7	0
PR3	0	8	0
PRI1	0	0	9
PRI2	0	0	10

PRI3	0	0	11

GAMMA			
	PS	PR	PRI
TR	12	13	14
PHI			
	PS	PR	PRI
PS	0		
PR	15	0	
PRI	16	17	0
PSI			
	TR		
	18		
THETA-EPS			
	TR1	TR2	TR3
	19	20	21

THETA-DELTA									
	PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
	22	23	24	25	26	27	28	29	30

Number of Iterations = 11

LISREL Estimates (Weighted Least Squares)

LAMBDA-Y	
	TR
TR1	0.917

TR2	0.955
	(0.022)
	43.035
TR3	0.828
	(0.026)
	31.585

LAMBDA-X			
	PS	PR	PRI
PS1	0.741	-	-
	(0.033)		
	22.446		
PS2	0.850		
	(0.026)		
	32.107		
PS3	0.869	-	-
	(0.021)		
	40.753		
PR1	-	0.889	-
		(0.025)	
		35.549	
PR2	-	0.960	-
		(0.015)	
		62.532	
PR3	-	0.956	
		(0.017)	
		57.426	

PRI1	-	-	0.968
			(0.011)
			88.903
PRI2	-	-	0.992
			(0.010)
			103.852
PRI3	-	-	0.993
			0.011
			94.592

GAMMA			
	PS	PR	PRI
TR	0.039	0.472	0.466
	(0.055)	(0.062)	(0.053)
	0.703	7.632	8.724

Covariance Matrix of ETA and KSI

	TR	PS	PR	PRI
TR	1.000			
PS	0.667	1.000		
PR	0.777	0.720	1.000	
PRI	0.771	0.619	0.593	1.000

PHI			
	PS	PR	PRI
PS	1.000		
PR	0.720	1.000	

	(0.035)		
	20.398		
PRI	0.619	0.593	1.000
	(0.036)	(0.038)	
	16.974	15.584	
PSI			
	TR		
	0.248		
	(0.029)		
	8.561		

Squared Multiple Correlations for Structural Equations

TR
0.752

Squared Multiple Correlations for Reduced Form

TR
0.752

THETA-EPS		
TR1	TR2	TR3
0.159	0.088	0.315
(0.054)	(0.051)	(0.059)
2.925	1.727	5.326

Squared Multiple Correlations for Y - Variables

TR1	TR2	TR3
0.841	0.912	0.685

THETA-DELTA								
PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.451	0.277	0.244	0.210	0.183	0.087	0.064	0.015	0.013
(0.066)	(0.063)	(0.057)	(0.062)	(0.061)	(0.054)	(0.049)	(0.048)	(0.048)
6.885	4.422	4.258	3.362	2.987	1.607	1.312	0.321	0.272

Squared Multiple Correlations for X - Variables

PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.549	0.723	0.756	0.790	0.817	0.913	0.936	0.985	0.987

Goodness of Fit Statistics

Degrees of Freedom = 48

Minimum Fit Function Chi-Square = 97.315 (P = 0.000)

Estimated Non-centrality Parameter (NCP) = 49.315

90 Percent Confidence Interval for NCP = (24.955 ; 81.448)

Minimum Fit Function Value = 0.186

Population Discrepancy Function Value (F0) = 0.0943

90 Percent Confidence Interval for F0 = (0.0477 ; 0.156)

Root Mean Square Error of Approximation (RMSEA) = 0.0443

90 Percent Confidence Interval for RMSEA = (0.0315 ; 0.0570)

P-Value for Test of Close Fit (RMSEA < 0.05) = 0.758

Expected Cross-Validation Index (ECVI) = 0.301

90 Percent Confidence Interval for ECVI = (0.254 ; 0.362)

ECVI for Saturated Model = 0.298

ECVI for Independence Model = 16.070

Chi-Square for Independence Model with 66 Degrees of Freedom = 8380.660

Independence AIC = 8404.660

Model AIC = 157.315

Saturated AIC = 156.000

Independence CAIC = 8467.798

Model CAIC = 315.160

Saturated CAIC = 566.396

Normed Fit Index (NFI) = 0.988

Non-Normed Fit Index (NNFI) = 0.992

Parsimony Normed Fit Index (PNFI) = 0.719

Comparative Fit Index (CFI) = 0.994

Incremental Fit Index (IFI) = 0.994

Relative Fit Index (RFI) = 0.984

Critical N (CN) = 396.995

Root Mean Square Residual (RMR) = 0.115

Standardized RMR = 0.115

Goodness of Fit Index (GFI) = 0.993

Adjusted Goodness of Fit Index (AGFI) = 0.989

Parsimony Goodness of Fit Index (PGFI) = 0.611

Model 3 (n=1048)

Observed Variables

Willingness to Buy	Perceived Size	Perceived Reputation	Trust	Privacy	Risk Perception
WTB1	PS1	PR1	TR1	PRI1	RP1
WTB2	PS2	PR2	TR2	PRI2	RP2
WTB3	PS3	PR3	TR3	PRI3	RP3

Sample Size: 1048

Latent Variables: TR PS PR PRI

Relationships

TR1 = TR

TR2 = TR

TR3 = TR

PS1 = PS

PS2 = PS

PS3 = PS

PR1 = PR

PR2 = PR

PR3 = PR

PRI1 = PRI

PRI2 = PRI

PRI3 = PRI

TR = PS PR PRI

Correlation Matrix

	TR1	TR2	TR3	PS1	PS2	PS3
TR1	1.000					

TR2	0.698	1.000				
TR3	0.635	0.698	1.000			
PS1	0.362	0.319	0.329	1.000		
PS2	0.4449	0.376	0.358	0.599	1.000	
PS3	0.436	0.416	0.355	0.453	0.584	1.000
PR1	0.547	0.516	0.531	0.418	0.459	0.436
PR2	0.556	0.487	0.467	0.342	0.441	0.438
PR3	0.564	0.578	0.499	0.347	0.429	0.460
PRI1	0.521	0.651	0.455	0.247	0.289	0.381
PRI2	0.524	0.669	0.493	0.290	0.337	0.388
PRI3	0.543	0.659	0.492	0.304	0.317	0.397

Correlation Matrix

	PR1	PR2	PR3	PRI1	PRI2	PRI3
PR1	1.000					
PR2	0.733	1.000				
PR3	0.720	0.784	1.000			
PRI1	0.359	0.367	0.417	1.000		
PRI2	0.364	0.350	0.408	0.904	1.000	
PRI3	0.377	0.365	0.420	0.888	0.912	1.000

Parameter Specifications

LAMBDA-Y	
TR	
TR1	0
TR2	1
TR3	2

LAMBDA-X			
	PS	PR	PRI
PS1	3	0	0
PS2	4	0	0
PS3	5	0	0
PR1	0	6	0
PR2	0	7	0
PR3	0	8	0
PRI1	0	0	9
PRI2	0	0	10
PRI3	0	0	11

GAMMA			
	PS	PR	PRI
TR	12	13	14
PHI			
	PS	PR	PRI
PS	0		
PR	15	0	
PRI	16	17	0
PSI			
	TR		
	18		
THETA-EPS			
	TR1	TR2	TR3
	19	20	21

THETA-DELTA									
	PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
	22	23	24	25	26	27	28	29	30

Number of Iterations = 9

LISREL Estimates (Weighted Least Squares)

LAMBDA-Y	
	TR
TR1	0.880
TR2	0.949
	(0.020)
	37.433
TR3	0.800
	(0.022)
	35.687

LAMBDA-X			
	PS	PR	PRI
PS1	0.689	-	-
	(0.029)		
	24.045		
PS2	0.832		
	(0.021)		
	38.741		
PS3	0.774	-	-

	(0.020)		
	38.342		
PR1	-	0.887	-
		(0.018)	
		49.869	
PR2	-	0.887	-
		(0.017)	
		51.961	
PR3	-	0.936	
		(0.013)	
		72.935	
PRI1	-	-	0.952
			(0.010)
			99.615
PRI2	-	-	0.980
			(0.009)
			110.220
PRI3	-	-	0.982
			0.011
			90.696

GAMMA			
	PS	PR	PRI
TR	0.111	0.420	0.465
	(0.042)	(0.044)	(0.036)
	2.650	9.604	12.857

Covariance Matrix of ETA and KSI

	TR	PS	PR	PRI
TR	1.000			
PS	0.653	1.000		
PR	0.744	0.688	1.000	
PRI	0.749	0.546	0.534	1.000

PHI			
	PS	PR	PRI
PS	1.000		
PR	0.688	1.000	
	(0.029)		
	24.081		
PRI	0.546	0.534	1.000
	(0.031)	(0.030)	
	17.674	17.624	
PSI			
	TR		
	0.268		
	(0.025)		
	10.735		

Squared Multiple Correlations for Structural Equations

TR
0.732

Squared Multiple Correlations for Reduced Form

TR
0.732

THETA-EPS		
TR1	TR2	TR3
0.226	0.099	0.361
(0.042)	(0.038)	(0.044)
5.401	2.621	8.202

Squared Multiple Correlations for Y - Variables

TR1	TR2	TR3
0.774	0.901	0.639

THETA-DELTA								
PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.526	0.307	0.401	0.213	0.213	0.123	0.094	0.041	0.037
(0.050)	(0.047)	(0.044)	(0.044)	(0.043)	(0.039)	(0.036)	(0.035)	(0.038)
10.484	6.501	9.116	4.819	4.930	3.143	2.628	1.142	0.974

Squared Multiple Correlations for X - Variables

PS1	PS2	PS3	PR1	PR2	PR3	PRI1	PRI2	PRI3
0.474	0.693	0.599	0.787	0.787	0.877	0.906	0.959	0.963

Goodness of Fit Statistics

Degrees of Freedom = 48

Minimum Fit Function Chi-Square = 106.795 (P = 0.000)

Estimated Non-centrality Parameter (NCP) = 58.795

90 Percent Confidence Interval for NCP = (32.652 ; 92.671)

Minimum Fit Function Value = 0.102

Population Discrepancy Function Value (F0) = 0.0562

90 Percent Confidence Interval for F0 = (0.0312 ; 0.0885)

Root Mean Square Error of Approximation (RMSEA) = 0.0342

90 Percent Confidence Interval for RMSEA = (0.0255 ; 0.0429)

P-Value for Test of Close Fit (RMSEA < 0.05) = 0.999

Expected Cross-Validation Index (ECVI) = 0.159

90 Percent Confidence Interval for ECVI = (0.134 ; 0.192)

ECVI for Saturated Model = 0.149

ECVI for Independence Model = 8.184

Chi-Square for Independence Model with 66 Degrees of Freedom = 8544.581

Independence AIC = 8568.581

Model AIC = 166.795

Saturated AIC = 156.000

Independence CAIC = 8640.037

Model CAIC = 345.434

Saturated CAIC = 620.462

Normed Fit Index (NFI) = 0.988

Non-Normed Fit Index (NNFI) = 0.990

Parsimony Normed Fit Index (PNFI) = 0.718

Comparative Fit Index (CFI) = 0.993

Incremental Fit Index (IFI) = 0.993

Relative Fit Index (RFI) = 0.983

Critical N (CN) = 723.374

Root Mean Square Residual (RMR) = 0.0726

Standardized RMR = 0.0726

Goodness of Fit Index (GFI) = 0.995

Adjusted Goodness of Fit Index (AGFI) = 0.992

Parsimony Goodness of Fit Index (PGFI) = 0.612

Banner screenshot

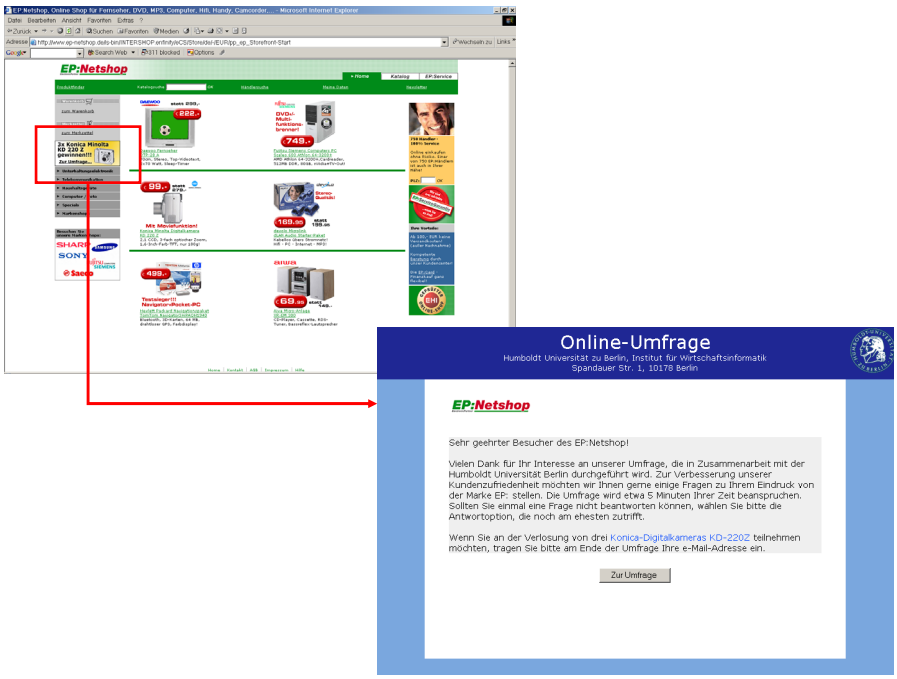

Figure 0-1: Screenshots of the banner leading to the survey

Appendix to Chapter 3

Survey results

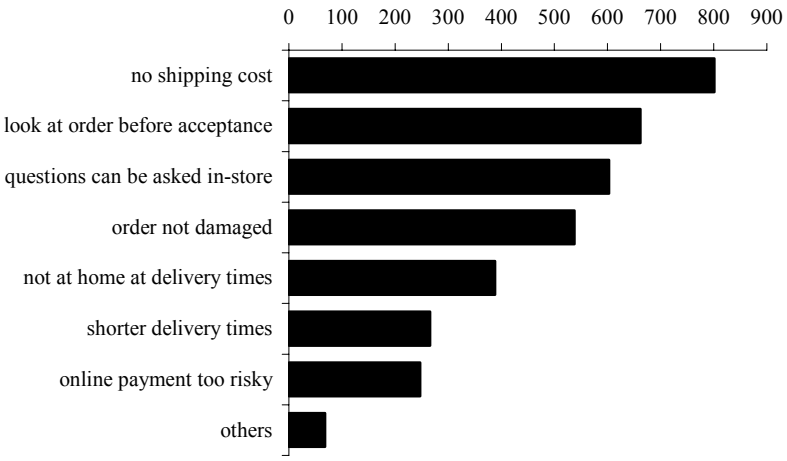

Figure 0-2: Frequency of answers to the question "if you have decided to pick up an online order at the retailer, what were the reasons?" (translated from German)

Customer and shop distribution

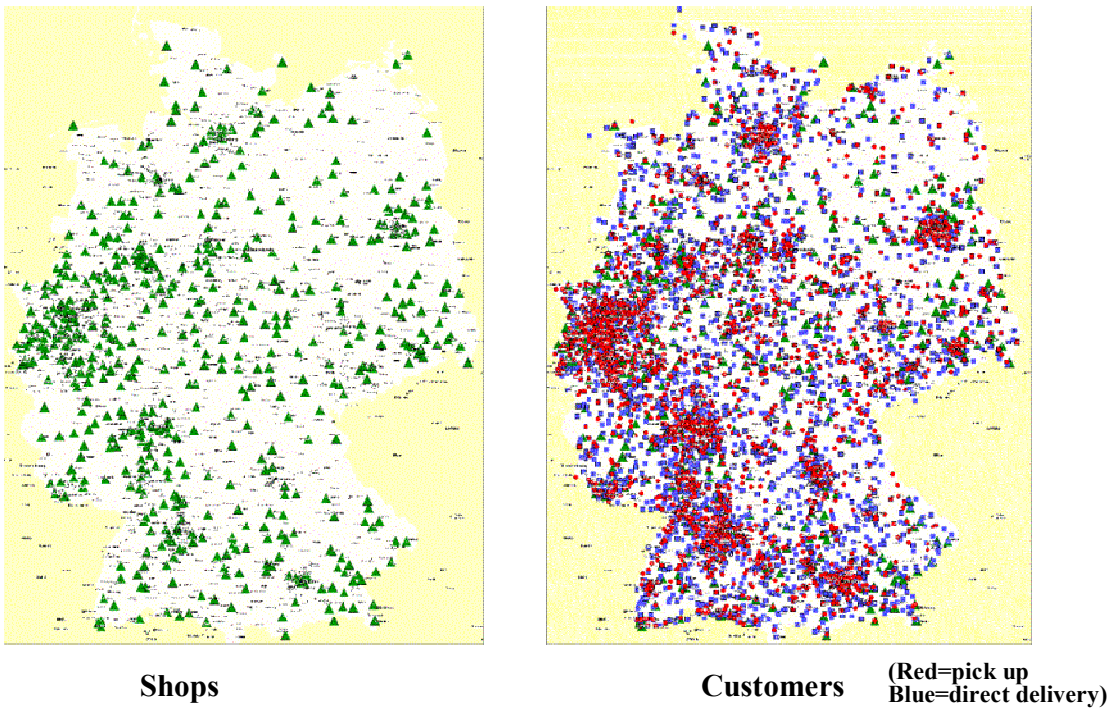

Figure 0-3: Shop and customer distribution of the multi-channel retailer

Analysis framework summary

*I: requires at least personal identification data, P: requires at least pseudonymous data, N: no modification required; N<P<I; all analyses are time-referenced

Category	Explanation	Required Data entities (cf. Table 3-2)	I*	P	N
Order Analyses					
Metrics					
Number of orders	Calculates the total number of orders in a given time frame	order_id, order_date			X
Mean number of purchases per customer	Calculates the arithmetic mean of the number of orders per customer	customer_id, order_id, order_date		X	
Mean transaction amount per customer	Calculates the arithmetic mean of the generated revenue per customer	customer_id, invoice_value, order_date		X	
Mean transaction amount per order	Calculates the arithmetic mean of the generated revenue per order	invoice_value, order_id, order_date			X
Mean margin per customer	Calculates the arithmetic mean of the margin per order	customer_id, invoice_value, order_id, cost, product_id, order_date		X	
Mean margin per order	Calculates the arithmetic mean of the margin per order	invoice_value, order_id, cost, product_id, order_date			X
Mean interpurchase time	Calculates the arithmetic mean of the time between two successive purchases of the same customer	customer_id, order_id, order_date		X	
Gini coefficient	Measures the concentration coefficient of generated revenue per customer	customer_id, invoice_value, order_id, order_date		X	
Analytics					
Recency distribution	Calculates classes of the number of customers that repeatedly puchased within the same time frame from their most recent visit and the present time	customer_id, order_id, order_date		X	
Purchase tenure distribution	Calculates classes of the number of customers that repeatedly puchased within the same time frame from their first visit and the present time	customer_id, order_id, order_date		X	
Frequency distribution	Calculates classes of the number of customers that incurred the same number of	customer_id, order_id, order_date		X	

	orders in a time frame				
Monetary value distribution	Calculates classes of the number of customers that generated the same range of order value in a time frame	customer_id, invoice_value, order_id, order_date		X	
Margin distribution	Calculates classes of the number of customers that generated the same range of profit margin in a time frame	customer_id, invoice_value, cost, product_id, order_id, order_date		X	
Recency for specific customer	Calculates the time between an individuals last purchase and the present time (individual specified by name and address)	customer_id, first_name, surname, address, order_id, order_date	X		
Frequency for specific customer	Calculates the number of order an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, order_id, order_date	X		
Monetary value for specific customer	Calculates the monetary value an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, invoice_value, order_id, order_date	X		
Margin for specific customer	Calculates the margin an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, invoice_value, cost, product_id, order_id, order_date	X		
Revenue contribution	Depicts the revenue contribution of classes of customers (Lorenz Curve)	customer_id, invoice_value, order_id, order_date		X	
Demographic Analyses					
Metrics					
Gender split	Calculates the ratio of female and male customers	customer_id, gender, order_date		X	
Mean revenue/gender	Calculates the arithmetic mean of the revenue generated by female and male customers	customer_id, invoice_value, gender, order_date		X	
Mean margin/gender	Calculates the arithmetic mean of the margin generated by female and male customers	customer_id, invoice_value, cost, product_id, gender, order_date		X	
Customer-Distance correlation	Measures the Person correlation between the number of customers - normalized with the population density in that zip code - with the zip code's distance to the next physical store	customer_id, customer_zip_code, geo_id, store_zip_code, longitude_zip_code, latitude_zip_code, order_date		X	
Analytics					
Mean revenue by age	Calculates classes of age and the	customer_id, date_of_birth,		X	

distribution	corresponding mean revenue per order	order_id, invoice_value, order_date			
Number of customers per location (zip code)	Calculates classes of locations and the corresponding number of customers	order_id, customer_id, customer_zip_code, order_date		X	
Number of transactions per location (zip code)	Calculates classes of locations and the corresponding number of transactions	order_id, customer_id, customer_zip_code, order_date			X
Revenue per location	Calculates classes of locations and the corresponding revenue	order_id, customer_id, invoice_value, customer_zip_code, order_date			X
Margin per location	Calculates classes of locations and the corresponding margin	order_id, customer_id, invoice_value, cost, product_id, customer_zip_code, order_date			X
Microgeographic details of customers in zip code area	Includes a variety of analytics describing microgeographic details of customers in a given zip code area	customer_id, order_id, address_id, customer_zip_code, geo_id, micro_id, detail_type, detail_value, order_date		X	
Microgeographic details for specific customer	Includes a variety of analytics describing microgeographic details of individual customers	customer_id, first_name, surname, address, geo_id, micro_id, detail_type, detail_value, order_date	X		
Product Analyses					
Metrics					
Mean number of products per order	Calculates the arithmetic mean of the number of products per order	order_id, product_id, order_date			X
Mean number of products in shopping cart	Calculates the arithmetic mean of the number of products put into the shopping cart per all customers who used the shopping cart	product_id, session_id, page_id, concept_name, access_time			X
Product Analytics					
Top-Selling Products	Returns a list of products that have been sold most frequently	order_id, product_id, product_name, order_date			X
Product associations	Returns a list or association rules between products with respective support and confidence values	order_id, product_id, product_name, order_date			X
Products per location	Calculates classes of locations and the respective top-selling product in that location	product_id, product_name, customer_id, order_id, customer_zip_code, order_date			X

Multi-Channel Service Analyses					
Metrics					
In-store payment ratio	Calculates the number of orders that were paid in-store per number of all transactions	order_id, payment_method, order_date			X
Online payment ratio	Calculates the number of orders that were paid online per number of all transactions	order_id, payment_method, order_date			X
Cash-on-delivery payment rate	Calculates the number of orders that were paid cash on delivery per number of all transactions	order_id, payment_method, order_date			X
In-store payment migration ratio	Calculates the number of repeat customers who changed payment preferences from online to in-store in at least one of the following transactions per number of all customers	customer_id, order_id, payment_method, order_date		X	
Online payment migration ratio	Calculates the number of repeat customers who changed payment preferences from in-store to online in at least one of the following transactions per number of all customers	customer_id, order_id, payment_method, order_date		X	
Pickup in-store ratio	Calculates the number of orders that were picked up in store per number of all transactions	order_id, delivery_type, order_date			X
Direct delivery ratio	Calculates the number of orders that were delivered directly per number of all transactions	order_id, delivery_type, order_date			X
In-store delivery migration ratio	Calculates the number of repeat customers who changed delivery preferences from online to in-store in at least one of the following transactions per number of all customers	customer_id, order_id, delivery_type, order_date		X	
Direct delivery migration ratio	Calculates the number of repeat customers who changed delivery preferences from in-store to online in at least one of the following transactions per number of all customers	customer_id, order_id, delivery_type, order_date		X	
Returns to stores ratio	Calculates the number of orders that were returned to physical stores per number of all transactions	customer_id, order_id, store_id, order_date		X	
Analytics					
Product weight and pick-up distribution	Calculates classes of the number of orders consisting of products within the same weight range and compares it with the number of pick-ups	order_id, product_id, product_weight, delivery_type, order_date			X

Product size and pick-up distribution	Calculates classes of orders consisting of products within the same size range and compares it with the number of pick-ups	customer_id, order_id, product_id, product_size, delivery_type, order_date			X
Revenues and pick-up distribution	Calculates classes of the number of orders consisting of products within the same revenue range and compares it with the number of pick-ups	customer_id, order_id, invoice_value, delivery_type, order_date			X
Returns from location distribution	Calculates the distribution of returns from locations (e.g zip codes)	customer_id, order_id, customer_zip_code, store_id, order_date			X
Returns/name and address	Calculates the distribution of individuals and respective number of returns	customer_id, order_id, first_name, surname, address, store_id, order_date	X		
Conversion Metrics					
Micro-Conversion Rates					
Look-to-click	Visitors who performed a product click-through/ visitors who saw a product impression	session_id, page_id, concept_name, access_time			X
Click-to-basket	Visitors who effected a basket placement/ visitors who performed a product click-through	session_id, page_id, concept_name, access_time			X
Basket-to-buy	Visitors who made a product purchase/ visitors who effected a basket placement	session_id, page_id, concept_name, access_time			X
Look-to-buy	Visitors who made a product purchase/ visitors who saw a product impression	session_id, page_id, concept_name, access_time			X
Life Cycle Metrics					
Reach	Suspects/ Whole Population	session_id, page_id, concept_name, access_time			X
Acquisition	Visitors who become active site investigators (Prospects/ Suspects)	session_id, page_id, concept_name, access_time			X
Conversion	Visitors who purchase (Customers/Prospects)	session_id, page_id, order_id, concept_name, access_time			X
Retention	Repeat Customers/ Customers	customer_id, order_id, order_date		X	
Life Cycle Interruption Metrics					
Abandonment	Visitors who filled shopping cart and	session_id, page_id,			X

	abandoned it/ active site investigators	concept_name, access_time			
Attrition	Customers who subsequently became customers elsewhere/ customers	customer_id, first_name, surname, address, order_date	X		
Churn	Attrited customers/ customers minus attrited customers	customer_id, first_name, surname, address, order_date	X		
Concept Conversion Rates					
Offline Conversion Rate	Visitors who accessed at least one page of the offline concept/ all visitors	session_id, page_id, concept_name, access_time			X
Store Locator visits	Visitors who access the store locator at least once/ all visitors	session_id, page_id, concept_name, access_time			X
Store Locator exits	Visitors who exited the site via the store locator/ all visitors	session_id, page_id, concept_name, access_time			X
Session Analyses					
Session clusters of all sessions	Calculates k-means session clusters of all sessions	session_id, page_id, concept_name, access_time			X
Session clusters of purchase sessions	Calculates k-means session clusters of all purchase sessions	session_id, page_id, concept_name, order_id, access_time			X
Session clusters of purchase sessions with direct delivery preference	Calculates k-means session clusters of all purchase sessions with direct delivery preferences	session_id, page_id, concept_name, order_id, delivery_type, access_time			X
Session clusters of purchase sessions with pick-up preferences	Calculates k-means session clusters of all purchase sessions with pick-up in store preferences	session_id, page_id, concept_name, order_id, delivery_type, access_time			X
Sessions/location distribution	Matches IP-addresses with geographic location	IP_address, zip_code, geo_id, micro_id, detail_type, detail_value,	X		
User typologies	Template matching with WUM mining language				
Further Web analyses					
Stickiness	Total amount of time spent viewing all pages by total number of unique users. Can be	session_id, page_id, concept_name, order_id,			X

	applied to entire sites or sections of sites.	delivery_type, access_time			
Slipperiness	1-Stickiness	session_id, page_id, concept_name, order_id, delivery_type, access_time			X
Focus	Number of pages visited in a given content section per total number of pages in the section.	session_id, page_id, concept_name, access_time			X
Velocity	Mean time users need to pass from one phase of the purchase process to the next	session_id, page_id, concept_name, access_time			X
Visit frequency	Mean number of visits by number of unique users	session_id, page_id, concept_name, cookie_id, access_time		X	
Mean visit duration	Calculates the arithmetic mean of visitors' total amount of time spent viewing all pages in session by number of all visits	session_id, page_id, concept_name, access_time			X
Visit tenure	Calculates classes of the number of visitors that visited site within the same time frame from their first visit	session_id, page_id, concept_name, cookie_id, access_time		X	
Visit recency	Calculates classes of the number of visitors that visited the site within the same time frame from their last visit	session_id, page_id, concept_name, cookie_id, access_time		X	
Visit retention	Visitors/repeat visitors	session_id, page_id, concept_name, cookie_id, access_time		X	
Required clicks to first purchase	Calculates classes of the number of visitors that required the same number of clicks to their first purchase	Customer_id, session_id, page_id, concept_name, order_id, access_time		X	
Required clicks to repeat purchase	Calculates classes of the number of customers that required the same number of clicks to their first repeat purchase	session_id, page_id, concept_name, order_id, access_time			X
Marketing Analyses					
Order value of referrals	Calculates the distribution of revenue generated through referrers	session_id, page_id, concept_name, order_id, access_time, invoice_value			X
Contest/game participation	Calculates the number of participants in contests/games	session_id, page_id, concept_name, order_id, access_time			X

Table 0-3: Analysis framework summary

Appendix to Chapter 6

Experimental workflow

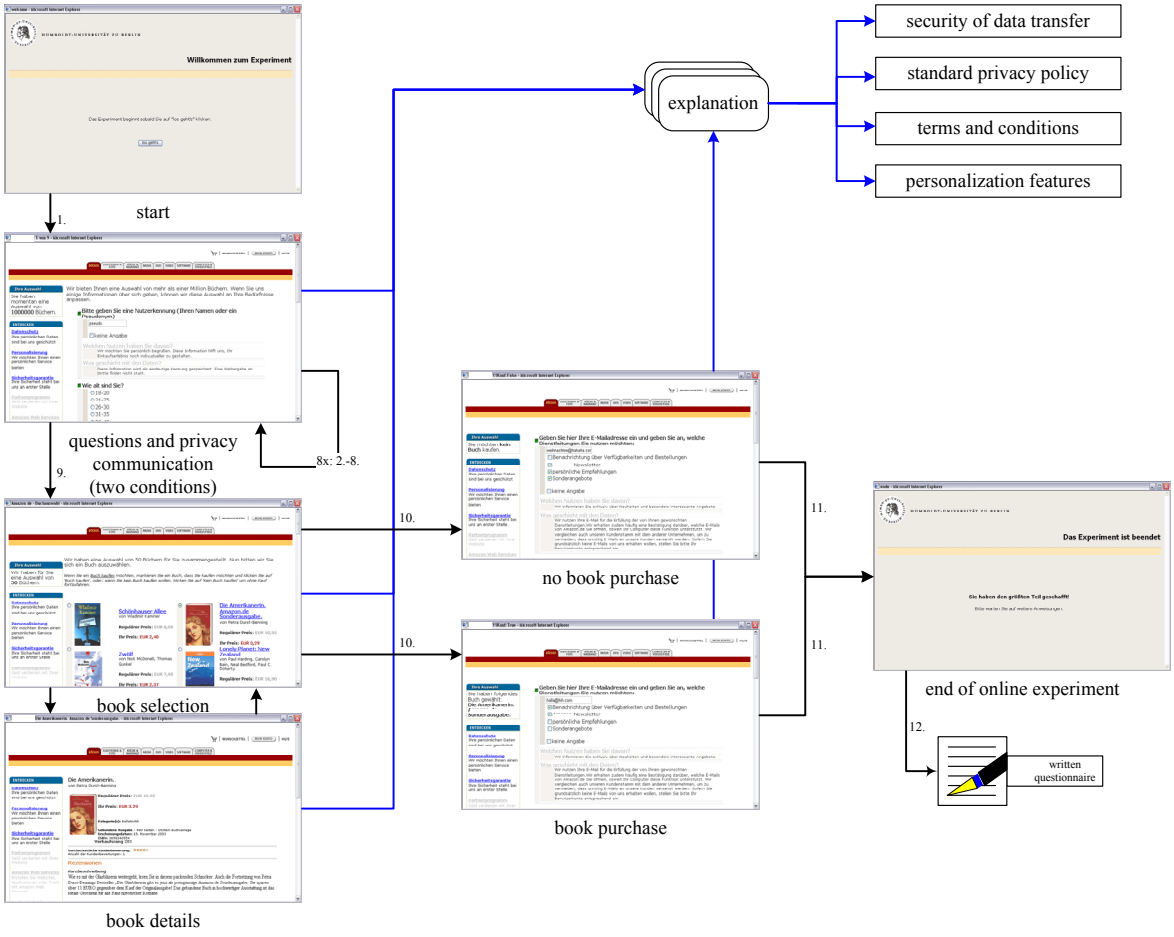

Figure 0-4: Workflow of the experimental procedure

Student briefing

Liebe Teilnehmerin, lieber Teilnehmer,

vielen Dank im Voraus für Ihre Teilnahme an diesem Experiment. Sie nehmen an einer Untersuchung teil, die von der Firma <Buchhändlername> und dem Lehrstuhl für Wirtschaftsinformatik der wirtschaftswissenschaftlichen Fakultät der Humboldt Universität zu Berlin durchgeführt wird.

Bitte vergewissern Sie sich, dass Sie Ihren Ausweis (bzw. ein Identifikationsdokument) dabei haben, sonst können Sie leider nicht am Experiment teilnehmen. Für den Fall, dass Sie ein Buch kaufen möchten, ist es unbedingt notwendig, dass Sie Kreditkarte bzw. Bankinformationen (z.B. Bankkarte) mitgebracht haben. Ohne diese Dokumente ist ein Buchkauf nicht möglich.

Ziel dieser Untersuchung ist, auf der Site von <Buchhändlername> einige Fragen eines Suchagenten zu beantworten. Nachdem Sie Ihre Interessen und Präferenzen angegeben haben, gibt Ihnen der Agent Buchempfehlungen. Wenn Sie eines der empfohlenen Bücher interessiert, können Sie es mit einem einmaligen Rabatt in Höhe von 70% erwerben.

Für die Beantwortung der Fragen haben Sie ca. 30 Minuten Zeit. Wenn Sie schon vorher fertig sind, bleiben Sie bitte an Ihrem Platz sitzen, bis auch die anderen Teilnehmer so weit sind. Sie haben genügend Zeit, die Fragen in Ruhe und aufmerksam zu beantworten. Damit unsere Untersuchung Erfolg hat, möchten wir Sie bitten, die Fragen des Suchagenten wahrheitsgemäß zu beantworten. Bitte beantworten Sie Fragen besser gar nicht, bevor Sie falsche Angaben machen.

Im Anschluss an das Kaufangebot präsentieren wir Ihnen einen kurzen Fragebogen online. Dieser Fragebogen hat nichts mit dem vorhergehenden Teil des Experimentes zu tun. Ihre Antworten werden unabhängig von Ihren vorherigen Angaben anonym gespeichert und ausgewertet.

Das Ausfüllen des Fragebogens dauert nur wenige Minuten. Wenn Sie damit fertig sind, melden Sie sich bitte bei uns! Sie bekommen dann Ihren Gutschein ausgehändigt und erhalten weitere Informationen, wenn Sie ein Buch gekauft haben.

Im Anschluss an den schriftlichen Fragebogen folgt ein weiterer, kurzer Fragebogen in elektronischer Form, den wir Sie bitten zu beantworten.

Wir möchten Sie bitten, über Ihre Teilnahme bis zum Ende der Experimentphase am Mittwoch um 18.00 Uhr Stillschweigen zu bewahren, damit die Ergebnisse nicht verfälscht werden.

Falls Sie noch Fragen haben oder während des Experiments noch Fragen auftreten sollten, so sprechen Sie uns einfach an. Wir werden Ihnen weiterhelfen.

Vielen Dank für Ihre Aufmerksamkeit und viel Spaß beim Experiment!

Questions in the experiment (with explanations)

Bitte geben Sie eine Nutzerkennung (Ihren Namen oder ein Pseudonym) an:_____

Welchen Nutzen haben Sie davon?

Wir möchten Sie persönlich begrüßen. Diese Information hilft uns, Ihr Einkaufserlebnis noch individueller zu gestalten.

Was geschieht mit den Daten?

Diese Information wird als eindeutige Kennung gespeichert. Eine Weitergabe an Dritte findet nicht statt.

Wie alt sind Sie?

○ 18-20

○ 21-25

○ 26-30

○ 31-35

○ 36-40

○ 41-50

○ 51-60

○ >60

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir suchen für Sie Informationen und Produkte heraus, die maßgeschneidert sind auf Ihr Alter. Diese Information hilft uns, Ihr Einkaufserlebnis individuell zu gestalten.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

In welchem Beruf/Studienfach sind Sie tätig? _____

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir suchen zu Ihrem Beruf passende Bücher und Informationen heraus.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Was sind Ihre Hobbys?

☐ Sport

☐ Musik

☐ Modellbau

☐ Computer

☐ Weitere_____

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Wir zeigen Ihnen, welche Bücher sie wirklich interessieren. Ihre Hobbys sind dabei ein wichtiges Kriterium.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Wir möchten gerne Cookies zur Aufzeichnung der Reihenfolge Ihres Aufrufs unserer Internetseiten (Clickstream) speichern. Sind Sie damit einverstanden?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir wollen Ihnen in Zukunft einen noch besseren Onlineauftritt bieten, der optimiert ist für Ihren Bildschirm und für Ihren Browser. Wenn Sie keine Cookies verwenden, sind Sie nicht in der Lage, so wichtige Features wie 1-Click®-Kaufen und "Neu für Sie" zu nutzen.

Was geschieht mit den Daten?

Ihre persönlichen Informationen verbleiben anonym. Ihr Einverständnis, Cookies zu akzeptieren, ermöglicht uns, unsere Site zu verbessern und Produkte und Dienstleistungen besser zu präsentieren. Weiterhin sind wir daran interessiert, Ihr wiederholtes Navigationsverhalten zu analysieren. Sie können Ihre Entscheidung, Cookies zu akzeptieren, jederzeit revidieren.

Geben Sie bitte Ihren Lieblingsautor oder -buchtitel ein: _____

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Wir berücksichtigen bei der Auswahl Ihren Lieblingsautoren und weitere, passende Bücher.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Worauf achten Sie beim Buchkauf mehr: Angaben auf dem Buchrücken oder Name des Autors?

○ Angaben auf dem Buchrücken (Kurzbeschreibung)

○ Autor

○ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherselektion wird dadurch noch besser. Autor und Inhalt werden unterschiedlich

stark berücksichtigt und der Bücherkauf wird noch einfacher für Sie.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Bitte teilen Sie uns mit, welche Bereiche Sie am meisten interessieren

☐ Antiquarische Bücher

☐ Belletristik

☐ Business & Karriere

☐ Börse & Geld

☐ Computer & Internet

☐ E-Books

☐ Fachbücher

☐ Film

☐ Kultur & Comics

☐ Geist & Wissen

☐ Hörbücher

☐ Kinder- & Jugendbücher

☐ Kochen & Lifestyle

☐ Krimis & Thriller

☐ Lernen & Nachschlagen

☐ Musiknoten

☐ Naturwissenschaften & Technik

☐ Politik

☐ Biografien & Geschichte

☐ Ratgeber

☐ Reise & Sport

☐ Religion & Esoterik

☐ Science Fiction

☐ Fantasy & Horror

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Wir können Ihnen eine bessere Auswahl empfehlen, wenn Sie uns mitteilen, welche Bereiche Sie am meisten interessieren.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welche Arten von Büchern kaufen Sie besonders häufig?

Hinweis: Bitte bilden Sie eine Rangordnung. (1 entspricht dabei dem am häufigsten gekauften Typ Buch)

__ Roman

__ Sachbuch

__ Fachbuch

__ Biographien

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Wir sortieren Bücher für Sie nach Ihren Interessen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Möchten Sie ein Buch als Ergänzung zu Ihren bisherigen Vorlieben kaufen, oder etwas komplett anderes?

○ Ergänzung zu meinen Vorlieben

○ etwas komplett anderes

○ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information hilft uns, Bücher nach Ihren Wünschen zu selektieren.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Leihen Sie Bücher?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Ihre Lesegewohnheit helfen uns, Sie noch besser kennenzulernen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welche Art Urlaubsreisen unternehmen Sie besonders gerne?

☐ Städtereisen

☐ Badeurlaube

☐ Familienreisen

☐ Erlebnisurlaube

☐ weitere

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Wir bieten Ihnen Bücher an, die sich rund um das Reisen drehen, und genau auf Sie zugeschnitten sind.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Reisen Sie eher pauschal, oder lieber auf eigene Faust?

○ Pauschal

○ Auf eigene Faust

○ gar nicht

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir bieten Ihnen Bücher an, die sich rund um das Reisen drehen, und genau auf Sie zugeschnitten sind.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Lernen Sie die Grundzüge einer Sprache, bevor Sie in ein anderes Land fahren?

○ ja

○ nein

○ keine Angabe

Sprachen:_____

Welchen Nutzen haben Sie davon?

Wir empfehlen Ihnen Bücher, die sich mit Fremdsprachen beschäftigen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Autoren welcher Nationalität bevorzugen Sie?

☐ Deutsch

☐ Englisch

☐ Französisch

☐ Spanisch

☐ Amerikanisch

☐ Andere

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information macht es uns möglich Ihre Vorlieben noch besser zu berücksichtigen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welche politische Richtung präferieren Sie bei Büchern?

linksextrem – rechtsextrem (Skala von 1-7)

○ politisch uninteressiert

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir suchen Bücher für Sie heraus, die Ihren politischen Interessen entsprechen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Mögen Sie lieber Marx oder Macchiavelli?

○ Marx

○ Macchiavelli

○ kenne ich nicht

○ mag beide

○ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information hilft uns, die wirklich interessanten Politik-Bücher für Sie noch genauer herauszufiltern.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Besitzen Sie bereits Bücher zu Gesundheitsthemen?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir suchen Bücher für Sie heraus, die Ihren Interessen wirklich entsprechen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert.

Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Zu welchen Gesundheitsthemen suchen Sie Antworten?

☐ Allergien

☐ Erkältungen

☐ Hautkrankheiten

☐ Chronische Krankheiten

☐ Geschlechtskrankheiten

☐ Weitere

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Bücher aus dem Gesundheitsbereich werden für Sie vorselektiert. Sie verbessern so unsere Empfehlungen und müssen sich nicht mit unnötigem Suchen aufhalten.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Sind Sie an Büchern zur Selbstmedikation interessiert?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir können Ihnen noch bessere Empfehlungen geben.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht

statt.

Interessieren Sie sich auch für alternative Heilmethoden?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welche Glaubensrichtung interessiert Sie bei Büchern im Bereich Religion?

○ Buddhismus

○ Christentum

○ Hinduismus

○ Islam

○ Judentum

○ Andere

○ keine Angabe

Welchen Nutzen haben Sie davon?

Wir suchen Bücher aus dem Bereich Religion für Sie heraus, die Ihren Interessen entsprechen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welcher Religion gehören Sie an?

○ Buddhismus

○ Christentum

○ Hinduismus

○ Islam

○ Judentum

○ weitere: _____

○ religionslos

○ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl zum Thema Religion passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Mögen Sie Liebesromane?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Sie finden dadurch schnell zu Ihrem Lieblingsroman.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welche Art erotischer Bücher mögen Sie?

☐ Mann/Frau

☐ Mann/Mann

☐ Frau/Frau

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Weshalb kaufen Sie Bücher?

☐ Unterhaltung

☐ Fortbildung

☐ Karriereförderung

☐ Zeitvertreib

☐ Nachschlagewerk

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Welchen Erzähltyp mögen Sie bei Büchern am liebsten?

☐ Kurzgeschichten

☐ Fabeln

☐ Dialogerzählungen

☐ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information hilft uns, den Erzähltyp bei Ihren Empfehlungen zu berücksichtigen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Mögen Sie lieber Bücher mit Einband oder Taschenbücher?

○ Bücher mit Einband

○ Taschenbücher

○ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Wieviele Bücher lesen Sie im Jahr?

○ 1-3

○ 4-5

○ 6-10

○ 11-15

○ mehr

○ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information hilft uns, die Anzahl der Bücherempfehlungen für Sie besser abzustimmen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Wieviel Geld geben Sie üblicherweise pro Jahr für Buchkäufe aus?

○ 1-50

○ 51-100

○ 101-200

○ 201-500

○ 501-1000

○ mehr

○ keine Angabe

Welchen Nutzen haben Sie davon?

Diese Information hilft uns, Ihnen nur Bücher anzubieten, die Sie sich auch leisten wollen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Wie hoch ist Ihr verfügbares Einkommen im Monat?

○ 0-500

○ 501-1000

○ 1001-1500

○ 1501-2000

○ mehr

○ keine Angabe

Welchen Nutzen haben Sie davon?

Die Bücherauswahl passt mit Hilfe dieser Angaben noch besser zu Ihnen.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Sind Sie an Schnäppchen interessiert?

○ ja

○ nein

○ keine Angabe

Welchen Nutzen haben Sie davon?

Schnäppchen aus Ihrem Interessenbereich werden für Sie herausgesucht.

Was geschieht mit den Daten?

Diese Information wird unter Ihrem Pseudonym gespeichert, aggregiert und analysiert. Die angezeigten Inhalte werden für Sie angepasst. Eine Weitergabe an Dritte findet nicht statt.

Questionnaire at the end of the experiment

Pseudonym oder Name (aus Experiment): _____

Hatten Sie das Gefühl, Ihre Daten sind bei <Buchhändlername> gut aufgehoben?

☐ stimmt gar nicht

☐ stimmt eher nicht

☐ teils-teils

☐ stimmt überwiegend

☐ stimmt völlig

☐ weiß nicht

Hatten Sie das Gefühl, Ihre Angaben haben <Buchhändlername> geholfen, interessante Bücher für Sie zu selektieren?

☐ stimmt gar nicht

☐ stimmt eher nicht

☐ teils-teils

☐ stimmt überwiegend

☐ stimmt völlig

☐ weiß nicht

Haben Sie den Nutzen Ihrer Datenangaben nachvollziehen können?

☐ gar nicht

☐ eher nicht

☐ teils-teils

☐ überwiegend

☐ völlig

☐ weiß nicht

<Buchhändlername> geht verantwortungsvoll mit meinen übermittelten Daten um:

☐ stimmt gar nicht

☐ stimmt eher nicht

☐ teils-teils

☐ stimmt überwiegend

☐ stimmt völlig

☐ weiß nicht

Datenschutz hat Priorität bei <Buchhändlername>:

☐ stimmt gar nicht

☐ stimmt eher nicht

☐ teils-teils

☐ stimmt überwiegend

☐ stimmt völlig

☐ weiß nicht

Weitere Kommentare:

Empfangene Unterstützung und Hilfe durch Kollegen

- Professor Kobsa gab wesentliche Beiträge und Ideen für Kapitel 5 und 6.
- Bertolt Meyer lieferte die methodische Grundlage für die Entwicklung und Auswertung des LISREL-Strukturgleichungsmodells in Kapitel 2.
- Sören Preibusch realisierte die Implementierung des datenschutzwahrenden Analysetools in Kapitel 4.
- Professor Spiliopoulou und Carsten Pohle führten die Auswertung der Daten in Absatz 3.7 durch.
- In Kapitel 2 sind Kommentare des Reviewprozesses durch Gutachter der Zeitschrift Information Systems Research, sowie der Conference of Electronic Commerce (ICEC 03) eingeflossen.
- In Kapitel 3 sind Kommentare des Reviewprozesses durch Gutachter der Multikonferenz Wirtschaftsinformatik 2004 und 2002, der IBM eBusiness Conference 2003, des ACM WebKDD Workshop 2003, der Conference on Electronic Commerce and Web Technologies (EC-Web 03), sowie der Zeitschrift Praxis der Wirtschaftsinformatik eingeflossen.
- In Kapitel 4 sind Kommentare durch Gutachter der IEEE Conference on Electronic Commerce (CEC04), sowie des IEEE Workshops on Privacy, Security, and Data Mining 2001 eingeflossen.
- In Kapitel 5 sind Kommentare durch Gutachter des CHI-2003 Workshop "Designing Personalized User Experiences for eCommerce: Theory, Methods, and Research" eingeflossen.
- In Kapitel 6 sind Kommentare durch Gutachter des Privacy Enhancing Technologies Workshop (PET 2004) eingeflossen.

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Berlin, den 2. Juli 2005

Maximilian Teltzrow

Eidesstattliche Erklärung

Hiermit erkläre ich, Maximilian Teltzrow, dass ich mich bisher noch an keiner Institution einem Doktorexamen unterzogen habe. Ferner wurde die Dissertation bisher an noch keiner anderen Fakultät vorgelegt.

Berlin, den 2. Juli 2005

Maximilian Teltzrow